Representation Learning for Behavioral Analysis of Complex Games

Shin Oblander*
Columbia University
Preliminary draft; please do not circulate

Current as of July 5, 2023; click here for the most recent version

Abstract

Many phenomena studied in marketing and economics are analyzed through the lens of non-cooperative games. Researchers often study empirical behavior of agents in simple games conducted in the lab, but research on complex real-world competitive settings with extensive action spaces and intricate payoff structures is rare due to methodological challenges and data limitations. To bridge this gap, I develop a novel neural network architecture that enables behavioral analysis of complex games by estimating a game's payoff structure (e.g., win probabilities between pairs of actions) while simultaneously mapping agent actions to a lower-dimensional latent space. I structure the neural network to enforce that the latent space encodes strategic similarities between actions in a smooth, linear manner. I apply my method to analyze a unique dataset of over 11 million matches played in a competitive video game with a large array of actions and complex strategic interactions. I find that players select actions that counterfactually would have performed better against recent opponents, demonstrating model-based reasoning. Still, players overrely on simple heuristics relative to model-based reasoning to an extent that is similar to findings reported in lab settings. I find that noisy and biased decision-making leads to frequent selection of suboptimal actions, which corresponds to lower player engagement. This demonstrates the limits of player sophistication when making complex competitive decisions and suggests that platforms hosting competitions may benefit from interventions that enable players to improve their decision-making.

1 Introduction

Many decision problems studied in marketing and economics involve the behavior of agents in *non-cooperative games*, where outcomes depend on multiple agents acting independently. For instance, when retailers decide what assortment of products to stock in their stores, their sales will depend on

*I am deeply thankful to Pokemon Showdown for the data and to Asim Ansari, Oded Netzer, and Olivier Toubia for their guidance. I also thank Christophe Van den Bulte, Daniel McCarthy, Donald Lehmann, Eric Bradlow, Eric Park, Eva Ascarza, Hengyu Kuang, Hortense Fong, Lan Luo (Columbia), Mark Dean, Matteo Alleman, Michael Woodford, Miklos Sarvary, Peter Fader, Robert Meyer, Ryan Dew, Silvio Ravaioli, Susannah Scanlan, Xinyu Wei, and participants in the Quantitative Marketing Lab and the Cognition and Decision Lab at Columbia for their helpful comments.

the products that other competing retailers choose to stock. When advertisers design commercials to air in prime-time TV advertising slots (e.g., the Super Bowl), they must consider which brands their competitors may advertise and what messaging they are likely to use. When sports teams consider rookie players to draft onto their team, they consider which players other teams are likely to draft. Understanding real-life agent behavior in such settings is critical for both academic researchers and industry practitioners. Researchers in behavioral economics and cognitive sciences wish to understand the processes humans and organizations use to make strategic decisions and when and how they deviate from normative optimality. Meanwhile, the stakeholders making such decisions wish to understand how to make normative improvements in their decision-making. Similarly, intermediaries that facilitate competition, such as sports leagues and two-sided platforms, wish to understand how to structure competitions and the decision-making environment to result in higher profit or social welfare.

In the classic rational agent model, where all players have complete information and unlimited ability to calculate the optimal action conditional on opponents' actions and anticipate the strategic behavior of other players, a Nash equilibrium arises where every player's action is optimal given the actions of all other players (Nash, 1951). Empirically, however, economic agents are known to often behave in a normatively suboptimal way, even in simple settings. Behavioral economists seek to understand when and how agents deviate from rationality, and in particular, the behavioral game theory literature investigates agent behavior in the context of non-cooperative games (e.g., Camerer and Ho, 2015). However, this research often studies simple games conducted in lab settings. In these games, the the action space is typically small, the payoff structure of the game is relatively easily understood (i.e., the payoff or reward a player receives as a function of their action and the actions of other players), and agents learn over the course of a relatively short lab session (e.g., one hour). It is unclear whether these empirical findings generalize to more complex real-world competitive environments.

More empirical work is needed to better understand the extent and implications of boundedly rational behavior in complex competitive settings, but such empirical work is challenging for two reasons. First, it is difficult to analyze the behavioral properties of player decisions in complex games based on empirical data. In observational settings, the payoff structure is usually not known to the researcher a priori and so must be empirically estimated, which is difficult when working with large action spaces (e.g., millions of possible actions). Additionally, a large action space makes it difficult to infer what decision rules players are using based on observations of their action selections: while some actions correspond to clear decision rules (e.g., best response), it is extremely unlikely that a player chooses one of these actions exactly. For instance, even if players attempt to follow a normatively optimal decision rule, their calculation of the optimum may be inherently noisy or imprecise, leading to the selected action being only approximately optimal (Sims, 2003). Thus, to measure the extent to which players approximately follow different decision rules, it is necessary to measure how "close" actions are to each other.

Second, it is difficult to obtain appropriate data for behavioral analysis of complex games: a large

number of observations is needed to analyze such games due to their large action spaces and complex payoff structures. This makes conducting games in the lab prohibitively expensive. Likewise, many observational empirical settings have limited sample sizes, and not all variables relevant to the agents' decisions are observed. For instance, when sports teams draft rookies, they select from a large pool of rookie players based on reports of player attributes assessed by internal scouts. Such reports are generally not available to outside researchers, and the number of drafting decisions observed will be small relative to the complexity of the attribute space of rookies.

In this paper, I address these difficulties by developing a novel supervised representation learning framework to enable behavioral analysis of decision-making in a complex game and demonstrate my approach on a unique dataset. My supervised machine learning framework models how player actions translate into payoffs (i.e., win probabilities) and simultaneously maps the action space to a smooth representation. The learned representations allow the characterization of actions as linear combinations of other actions, enabling behavioral interpretations of player decisions. For instance, even if it is unlikely that a player chooses an action that *exactly* corresponds to a decision rule like best response, they may choose an action that is "close" to the best response in terms of its strategic properties. They may also choose an action that combines some strategic properties of multiple different heuristics or decision rules. The representations learned by my model linearly encode the strategic properties of actions, such that player decisions can be approximated as linear combinations of decision rules. This, in turn, allows me to analyze the behavioral properties of player behavior.

To address the problem of data availability, I collect a comprehensive and novel dataset of player decisions in a complex, competitive, player-versus-player game in the popular Pokemon video game franchise. The decision problems faced by e-sports (competitive video game) players mirror the structure and complexity of many real-world competitive settings studied by economists and marketers, with motivated and sophisticated decision-makers, but unlike many other settings, it is feasible to obtain large and high-quality data. I further discuss the merits of e-sports as an empirical setting in Section 3.1.

As such, my contribution is twofold. First, I contribute a novel representation learning framework for analyzing complex games. To ensure that the learned representation of the action space permits behavioral interpretations, I model the log-odds of one action winning over another as a bilinear form of their embeddings. This enforces that linear combinations of embeddings linearly combine the strategic attributes of their corresponding actions, and that actions close together in the embedding space share similar strategic properties. The bilinear constraint is analogous to the dot product between embeddings used in the word2vec model of word co-occurrence, which enforces that linear combinations of word embeddings will smoothly combine the contexts in which those words are likely to appear (Mikolov et al., 2013). I show that my model achieves comparable performance to a more standard feed-forward neural network architecture that does not constrain the embedding space to have a linear structure; thus, imposing this restriction yields greater interpretability of the action space and enables downstream behavioral analysis without sacrificing the expressive capacity of the neural network.

Second, I contribute new substantive findings about how agents make decisions in complex games. I find that players do not solely base their decisions on past directly observed actions and payoffs, but also are able to perform model-based reasoning (e.g., reasoning counterfactually about best responses) and incorporate external information into their decisions, even with such complex action spaces and payoff structures. However, the average decision weights placed on these factors are much smaller than the weights placed on factors directly observed by the player. This is true even for highly skilled players. Additionally, as is to be expected in such a complex setting, player decisions are noisy, with the vast majority of variance in decisions unexplained by the decision variables in my model. Thus, while players can reason strategically even in complex games, they make noisy decisions and overrely on simple heuristics. Still, the extent of noise in the decision process and relative weights placed on model-free vs. model-based factors are comparable to those documented in simpler lab settings previously studied in the literature (Camerer et al., 2002; Khaw et al., 2017).

I further investigate the normative implications of these findings for players and the managerial implications for the game platform in my empirical context. I show that players are adaptive in deciding whether to change actions from one period to the next (i.e., they tend to change actions when it is more likely to lead to a normative improvement), but are not adaptive in deciding what decision rules to consider when selecting an action. Players thus tend to place too much weight on simple but poorly adapted heuristics and too little weight on decision rules that are difficult to compute but would lead to normatively better outcomes. Additionally, the noise in player decisions is consequential, with high-error decisions leading to worse normative performance. The normative suboptimality of player decisions has implications for the game platform in my empirical setting. I find that when players select an action that performs worse than their previous action, they are likely to get discouraged and quit playing. This suggests that the noise and/or suboptimal decision weights in player decisions may put players at greater risk of dropping out of the game. The platform may therefore be able to improve player engagement and retention by introducing decision tools that can help players identify their "blind spots" and reduce noise in the decision-making process.

The rest of this manuscript proceeds as follows. In Section 2, I discuss relevant literature in behavioral economics, representation learning, and marketing of video games and discuss my contribution to each of these streams. In Section 3, I describe my empirical context and dataset. In Section 4, I present and discuss model-free descriptive patterns in the data. In Section 5, I present my supervised representation learning model, report its empirical performance, and discuss its generalizability to other empirical settings. In Section 6, I use the results of the representation learning model to analyze the behavioral properties of player decisions, illustrating results about how players decide whether to change their action from one time period to the next and how to select a new action conditional on making a change. In Section 7, I explore the normative and managerial implications of my findings about player behavior. In Section 8, I conclude and discuss how my methodology can be applied to analyze agent decisions in other complex competitive settings. In the Appendices, I provide more details about the data and methodology and robustness checks of the empirical results.

2 Connections to Literature

This work contributes to three areas of literature. First, I contribute a new perspective to the extensive literature in behavioral economics that seeks to understand how real humans deviate from the rational agent model used in classic models of game theory and consumer behavior. In non-competitive settings, extensive work documents how agents deviate systematically from full information rationality across many types of consequential decision problems. For instance, consumers often make suboptimal personal finance decisions at substantial personal cost (e.g., Gathergood et al., 2019). Normative suboptimality in decision-making is not limited to lay consumers: for instance, in the context of macroeconomic forecasting, even professional forecasters and commercial banks exhibit systematic deviations from full information rational expectations (Coibion and Gorodnichenko, 2015).

Deviations from the rational agent model are ubiquitous and consequential, and as such, much work seeks to model and understand these deviations from rationality (Barberis, 2013). For instance, Houser et al. (2004) and Liu and Ansari (2020) analyze what decision rules agents use in dynamic discrete choice tasks conducted in the lab and when their decisions deviate from normative optimality. Various mathematical models of "boundedly rational" decision-making have been proposed, formulating agent behavior as arising from imperfect cognitive processes involving bias and/or noise. One such class of models is models of limited attention, wherein agents have finite cognitive capacity for processing information, which they strategically allocate towards more important information (Matějka and McKay, 2015; Khaw et al., 2017; Gabaix, 2019). While many models of bounded rationality are theoretical or estimated on lab data with stylized tasks, some researchers have begun to empirically estimate models of bounded rationality on observational empirical data in consumer choice settings (Joo, 2022).

The aforementioned works study decision-making in non-competitive settings. Turning to the setting of non-cooperative games, researchers in behavioral game theory model how players select and update actions over time in competitive games, usually using lab experiments with simple games (Camerer and Ho, 1999; Camerer et al., 2002; Ho et al., 2007; Camerer and Ho, 2015). Many researhers model action selection in repeated games as a learning problem, wherein an agent learns the expected payoffs of different actions over time and updates their actions in future periods accordingly. Researchers often seek to understand the extent to which players engage in model-free learning, wherein they directly learn a mapping from actions to payoffs without accounting for the opponent's action and the payoff structure of the game, compared to model-based learning, wherein they account for the payoff structure of the game to reason counterfactually about what actions could have performed better against a given opponent's action and account for this information in future time periods. Relatedly, researchers in evolutionary economics model the equilibration of games as a process of evolution, often using simulation studies (Samuelson, 1998). Models of quantal response equilibria treat each agent as rational up to having some probability of making an error, with all agents aware of their own (and other players') error probabilities (McKelvey and Palfrey, 1995).

I contribute to these streams of literature by extending behavioral analysis of agent decisions in games to a complex and naturalistic empirical setting, going beyond lab studies with simple games and revealing new insights as to the dynamics of agent learning and action selection with highly complex decisions. Closely related is the work of Howard (2021), who uses observational data on chess players in timed games to test the extent to which players are able to optimally trade off time spent deciding on a move and the resulting quality of the move. In contrast, I study agent learning over time and infer the decision rules that agents use to select actions.

Second, I contribute to the representation learning literature on methods for extracting meaningful low-dimensional summaries of complex high-dimensional data that can be used for downstream inference tasks. Neural network models have revolutionized analysis of high-dimensional and unstructured data such as text and images by extracting information from complex raw data and mapping it to a *representation* (often referred to as an "embedding"). These representations encode the information in a simplified way that makes downstream tasks (e.g., prediction) more tractable (Bengio et al., 2013). This includes "supervised" settings, where the goal is to extract information from a set of input variables to predict a dependent variable (e.g., identifying what object is contained in an image; LeCun et al., 2015), and unsupervised or self-supervised settings, where the goal is to model the dependency structure among many variables (e.g., large language models of the co-occurrences of words; Vaswani et al., 2017). The representations learned by such models often allow for interpretability, such as in the "word2vec" model (Mikolov et al., 2013), wherein words are represented as points in a Euclidean space, and arithmetic in this embedding space captures the semantic relationships between words (e.g., the embedding for "king" is approximately equal to the embeddings of "queen" plus "man" minus "woman").

Recent work in marketing has developed representation learning methods to extract meaningful structures from complex data that can then be used to obtain managerial insights. For instance, Dew et al. (2022) use multiview representation learning to embed brand logos in a latent space that simultaneously captures information about the brand's personality and other attributes. Analysis of the resulting embedding space allows for inferences about the similarities between different brands' aesthetics and personalities, enabling "brand arithmetic" that can propose new brand attributes that interpolate between different reference brands. Burnap et al. (2023) embed images of product designs into a latent space that allows for generation of new product designs that combine features of other reference designs. Beyond analysis of the embedding space itself, researchers also use the learned representations to enable modeling of human behavior in contexts where the original data would be infeasible to model directly: for instance, Ruiz et al. (2020) model consumer choice over a large assortment of products by embedding products into a latent space and modeling consumer preferences and price sensitivities as a function of those embeddings, rather than modeling preferences over the products directly. My research continues in this vein, proposing a representation learning model for games involving large, complex action spaces, yielding embeddings of actions that can then be used for downstream analysis of player choices of actions in the game. This allows for modeling of player decisions by performing analysis on the estimated embeddings rather than the

actions directly.

Third, I contribute to the budding empirical literature on the marketing and economics of video games. Several studies in marketing and related fields have analyzed video games as an entertainment medium, analyzing usage patterns and identifying factors that keep players and spectators engaged (Nevskaya and Albuquerque, 2019; Huang et al., 2019; Simonov et al., 2022). Other studies have modeled the demand for video games and consoles, considering implications for pricing and firm strategy (Ishihara and Ching, 2019; Haviv et al., 2020). However, there is relatively little research so far that seeks to empirically understand *how* players play video games. Some computer science research aims to predict in-game player behavior to train AI players or design adaptive content in games (Hooshyar et al., 2018), but these models generally do not provide insight into the behavioral processes driving player decisions.

In contrast, I empirically analyze how video game players behave in competitive settings, modeling how players learn about and respond to the actions of their competitors, so as to better understand the behavioral processes of player learning and action selection and assess the normative implications of player decisions. This is similar in spirit to cognition research analyzing the memory encoding tactics used by advanced players of strategy games such as chess (Gobet and Simon, 1998). I also show evidence that the behavioral properties of player decisions are consequential for game platforms, who may be able to improve player engagement and retention through interventions that help prevent players from making poor decisions.

3 Data and Empirical Context

3.1 Empirical Context

In this paper, I empirically study the behavioral properties of agent decisions in the context of an e-sports league, i.e., a competitive video game league. E-sports are a compelling empirical context for studying strategic decision-making: the games played in e-sports competitions are often highly complex, yet many players are highly motivated and sophisticated adults who devote considerable time and energy to optimizing their gameplay. While the games are complex, they are also largely self-contained environments, such that the variables relevant to analyzing player decision-making are observable to the researcher. Additionally, e-sports have exploded in popularity over the past several years: for instance, in 2021, the popular multiplayer battle game Fortnite boasted over 80 million monthly active players, with spectators viewing over a billion hours of livestreams on platforms such as Twitch and YouTube.[1] Given this popularity, it is feasible to obtain large datasets that cover many players over a long time horizon (e.g., weeks/months rather than a single lab session).

All of these features make e-sports a compelling empirical testing ground for studying how agents make complex decisions. Furthermore, beyond being an examplar to obtain general insights about human decision-making, the behavioral patterns of video game players are substantively relevant to understand in their own right. Video gaming is a large and growing industry where understanding

[1] https://www.businessofapps.com/data/fortnite-statistics/

consumer behavior is essential for game platforms that wish to attract and retain players. I show in Section 7.2 that normative suboptimality in player decision-making has consequences for player retention, such that understanding the decision-making process is managerially important for game platforms that wish to keep players engaged.

Specifically, in this paper, I study player behavior in the Pokemon video game franchise. Pokemon is one of the world's most successful media franchises, with hundreds of millions of copies sold of its main line video games,[2] cumulatively garnering an estimated $105 billion in revenue as of 2021.[3] Though primarily known for the single-player adventure mode of its games, the Pokemon franchise has a sizeable competitive scene based on in-game player-versus-player battles, known as the Video Game Championships (VGC) league. Each year, The Pokemon Company hosts dozens of regional and international tournaments for its video games and trading card game, where players compete for prizes and an invite to international and world championships.[4] The 2022 world championship tournament featured thousands of skilled players, over $1 million in prizes, and garnered over 8.5 million views on official English- and Japanese-language livestreams on Twitch and YouTube.[5] Beyond these tournament events, players can play on-demand ranked matches online at any time through Pokemon's video games. On-demand ranked matches are highly popular, with hundreds of thousands of monthly active users in the official league.[6] Though often perceived as a game franchise oriented towards children, the competitive scene consists largely of adult players with many years of experience who devote considerable time and energy to the game: for instance, in the 2023 European international championships, 89% of VGC contestants played in the "Masters" (adult) division of the tournament.[7]

For this analysis, I focus on online on-demand ranked matches, where players do not know each others' team compositions ahead of time and are matched effectively at random with another player of similar skill. The data for my analysis is provided by the fan-made website Pokemon Showdown,[8] which provides an open source emulation of the player-versus-player battle system used in official Pokemon leagues. The rules and gameplay mechanics on the VGC format of Showdown are identical to official games, and the website is a major platform for competitive Pokemon gameplay.

In the VGC format, each player builds a team of six characters (Pokemon) before being matched with an opponent, such that a player must make their team build decisions before knowing who their specific opponent will be; as such, they must build their team based on their prior beliefs about about the distribution of actions that may be used by a random opponent. Once matched, the two players proceed to play a turn-based battle until one player has all of their characters defeated (or a player forfeits).[9] For this project, I seek to understand the decisions made by players

[2] https://www.statista.com/statistics/1072224/pokemon-unit-sales-worldwide/

[3] https://www.statista.com/statistics/1257650/media-franchises-revenue/

[4] https://www.pokemon.com/us/play-pokemon/pokemon-events/pokemon-tournaments/

[5] https://www.pokemon.com/us/play-pokemon/worlds/2022/about/. View counts retrieved from the official (English) Pokemon Twitch channel and (English and Japanese) Pokemon YouTube channels on October 11, 2022.

[6] https://www.reddit.com/r/VGC/comments/n2h218/data_viz_how_many_players_play_and_hit_master/

[7] https://www.pokemon.com/uk/play-pokemon/internationals/2023/europe/event-results/#pokemon-vgc

[8] https://pokemonshowdown.com/

[9] Specifically, after matching and seeing the six other characters brought by the opponent, each player selects a

before having any specific knowledge of their opponent, and so I focus on analyzing the team building decisions. Specifically, I model the match win/loss outcomes in reduced form as a function of the teams brought by each player, rather than analyzing the sequence of actions that occurs after two players have been matched. Given the focus on team building, in all subsequent exposition I will use the term "action" to refer to the set of 6 Pokemon characters chosen by a player.[10] An example action is given in Figure 1.

Notably, events that occur in one match do not carry over to the next: for example, characters do not "level up" from experience or retain state variables (e.g., damage taken) from earlier matches. Additionally, whereas team building can be extremely time-intensive in official Pokemon games (requiring many hours in the single-player mode of the game to obtain the desired characters), Pokemon Showdown allows players to simply specify their desired team. Thus, characters do not need to be "unlocked" over time, and all players can access all actions with no monetary cost or time investment. Accordingly, the decision problem is in principle static and homogeneous, and so any dynamic behavior will arise from player learning, and any differences in actions across players will arise from heterogeneity and stochasticity in player decision-making.

Even just focusing on the team building decision, this setting yields a massive action space with a complex payoff function. In the time period of the data that I analyze, there are over 700 unique characters that are legal to use in the game, resulting in effectively innumerable possible combinations of six characters. Characters have non-transitive relationships analogous to "rock-paper-scissors": for instance, grass type Pokemon usually have an advantage over water types, water types usually have an advantage over fire types, and fire types usually have an advantage over grass types. Characters can also have non-additive complementarities with each other (e.g., a character may only perform well when paired with a specific other character), further complicating the payoff function. Game outcomes are also highly stochastic, such that even after observing a game outcome of two actions, the player may still be highly uncertain as to the probability of success of one action over the other.

The result is that action selection is an incredibly complex decision—there are effectively an innumerable number of combinations of characters that could be chosen to comprise the team of six. Since a given action will tend to perform well against some opposing actions but not others, the normative fitness (i.e., overall performance of an action, or how normatively well-adapted it is) of an action depends on the distribution of actions being used by other players. This distribution is not directly observed to the player, though a player may gather outside information about the distribution by observing other players who livestream their games or consulting publicly available

subset of four characters to bring to the game. Once this selection is completed, a turn-based battle is initiated, where each player has two characters in play at a time.

[10] In addition to selecting the six characters to include on the team, each character has a number of attributes, such as numeric stats (e.g., attack and defense) and moves/attacks, which can be customized by the player to some extent. For simplicity, I do not include data about attribute customization in my analysis; as discussed in Appendix A, in practice most differences between teams are captured just by the assortment of six characters: there is not much variability in attributes for a given character. Still, it would be straightforward to extend the proposed model to include customizable attributes by simply including the attributes as further inputs to the model, albeit at greater computational cost and model complexity.

Figure 1: Example action (Pokemon team)

An example of a Pokemon team build. This team was used by player Eduardo Cunha to win the 2022 World Championships (https://www.pokemon.com/us/play-pokemon/worlds/2022/vgc-masters/#team-eduardo-cunha-portugal-masters-division-champion). A team consists of six characters ("species"), which must all be unique (i.e., no two characters on the team can be the same species). Some species, such as Calyrex, Thundurus, and Zacian in this example, have multiple "forms" with different strategic attributes, which I treat as separate characters for the purposes of analysis.

summary statistics.[11] Due to the non-transitive, non-additive strategic properties of characters, there are not clear dominant strategies, and simplifying heuristics are not guaranteed to perform well. Players may also have difficulty determining how actions map onto payoffs.

Lastly, a unique feature of the VGC format is that the rules of the game change every few months: The Pokemon Company periodically changes which characters are allowed to be used (and sometimes, which mechanics are allowed to be used within the match). This introduces a discrete shock to the player learning process that allows observation of multiple "initial conditions" and how players react to a sudden forced change of action.

In sum, I study a complex repeated decision problem with a combinatorially large action space, where the outcomes of a player's action is dependent on the actions of other agents. The distribution of actions used by opponents is not directly observed and may evolve over time, though players may consult external resources to obtain partial information about the distribution. Beyond the lack of full information, the large action space and complexity of the payoff structure of the game make it difficult for players to optimize their actions (since they in principle need to consider all possible pairs of actions). Nonetheless, the players are experienced and motivated, and so they attempt to approach the decision with a high degree of sophistication.

The decision problem mirrors the complexity of other substantive settings of interest to marketing and economics researchers (e.g., competing retailers choosing assortments of products to stock). Though the complexity of the setting makes empirical analysis difficult, the popularity of the game makes it feasible to obtain sufficiently large datasets to enable rich analyses. Furthermore, the variables relevant to player decisions (e.g., past win/loss outcomes and the population distribution of actions) are all observable to the researcher, and the actions are fully observed as well (as opposed to, e.g., sports team drafting where not all rookie attributes used to make drafting decisions will be observable to an outside researcher). It is all of these features make Pokemon VGC a compelling

[11] E.g., the website Pikalytics posts aggregated summary statistics about character usage on Pokemon Showdown, which are updated once a month. https://www.pikalytics.com/

Table 1: Summary of observed rulesets

Ruleset	Start date	Restricted characters	Dynamax
Series 8	3/1/2021	Up to 1	Yes
Series 9	4/16/2021	None	Yes
Series 10	7/9/2021	Up to 1	No
Series 11	10/23/2021	Up to 1	Yes
Series 12	1/1/2022	Up to 2	Yes

Note: "Start date" refers to the first full day that the given ruleset appears in the data. The "restricted characters" column indicates how many members of the team may be from a list of 30 powerful characters that are considered "restricted"; in the example in Figure 1, Calyrex and Zacian are considered restricted (the 2022 World Championships took place under Series 12 rules, where teams with up to 2 restricted characters were legal). Dynamax is a mechanic that can be used by each player once per match which, roughly speaking, makes a selected character more powerful for three turns; this does not directly change which characters are allowed.

empirical setting to study the behavioral properties of strategic decision-making. Additional institutional details about the Pokemon VGC game and the Pokemon Showdown platform are given in Appendix A.

3.2 Data

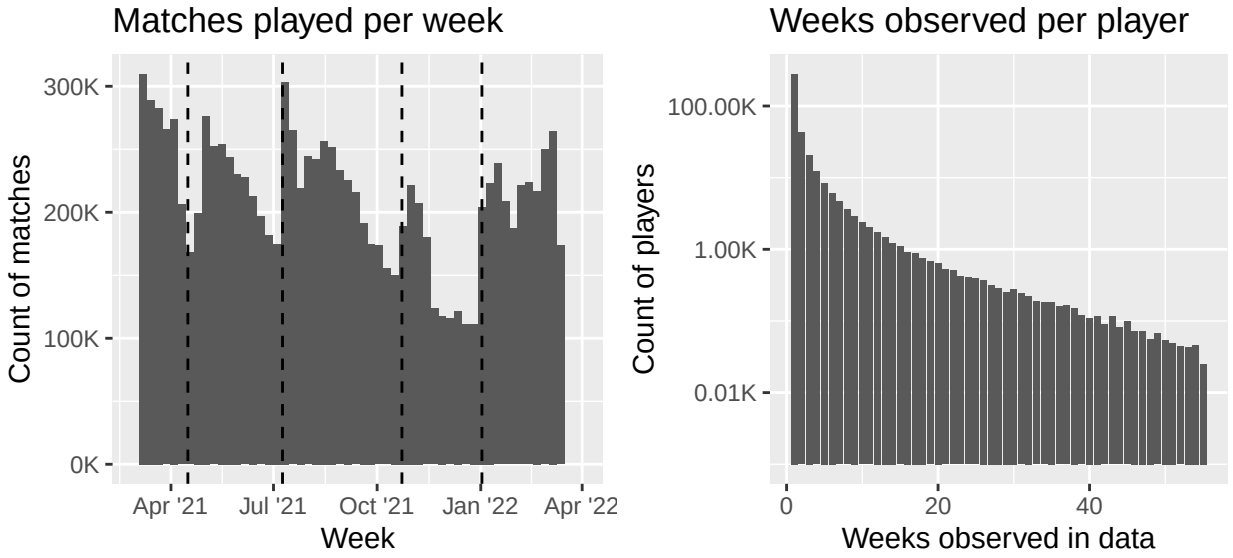
My analysis is based on approximately one year of data, covering all VGC format matches played on Pokemon Showdown between March 1, 2021 and March 17, 2022. In total, I observe 11,738,418 matches across 400,009 unique players and five rulesets (i.e., four rule changes). Table 1 summarizes the rulesets in effect over the data period.[12] For each match, I observe a unique (anonymized) identifier for the two players, each player's action (i.e., their set of six characters), a timestamp censored to the hourly level, and the winner. Thus, I observe 23,476,836 total player-action pairs (two per match). I observe 702 unique characters in the data, which are observed in 1,516,152 unique combinations, such that each action is observed an average of 15.5 times; the median number of observations per action is 4. This emphasizes the sheer size of the action space and sparsity of observations: 17.8% of actions in the data are only observed once. Evidently, given the size of the action space, it is nontrivial to estimate the probability of winning between pairs of actions, and thus not straightforward to assess behavioral and normative properties of player decisions.

Figure 2 shows some basic summaries of the data. Slightly over 200,000 matches are observed per week on average, with play activity usually heaviest in the first few weeks after a rule change. Most players only appear in a few weeks of data, but thousands of players consistently play over many weeks, allowing for longitudinal analysis of how these players learn and change their actions over time.

Though most of the 702 unique characters appear fairly rarely in the data, there is still substantial

[12]Here, the start date reflects the first full day that the ruleset appears in the Pokemon Showdown data, which is generally several weeks before the start date of the ruleset in the official Pokemon league; though The Pokemon Company announces new rulesets several weeks ahead of time, Pokemon Showdown implements new rules almost immediately after they are announced, such that players in my dataset do not have advance notice about rule changes.

Figure 2: Data summaries



Note: in the left panel, the vertical lines indicate dates of rule changes.

diversity in the team builds employed by players: for instance, there are 97 characters that appear on at least 1% of teams. Even selecting only from these relatively popular characters, there are still nearly a billion possible combinations of six characters. This indicates that the action distribution is not concentrated on a small set of dominant actions, and so a player who wishes to build a competitively viable team will need to take into account numerous possible actions that may be chosen by their opponent. I provide further details about the data and cleaning/pre-processing steps in Appendix B.

3.3 Measuring In-Game Skill

As mentioned above, after two players have matched, they participate in a turn-based game until one player wins. In my subsequent analyses, I model the game outcome in reduced form and do not consider the sequence of in-game actions. However, players may vary in their skill at the turn-based game component. This skill at in-game play is ancillary to the problem at hand, but it may be correlated with the player's skill at team selection. So, it is important to control for skill differences at in-game play to ensure that match outcomes due to player differences in in-game skill do not get falsely attributed to player action selection.

Accordingly, it is useful to have a measure of in-game skill that can be included in models of match outcomes to control for skill differences. Pokemon Showdown rates players using the Elo rating system (Elo, 1978), generally matching players with similar ratings together, but over half of the ratings are missing from the data due to inconsistent recording; additionally, Elo ratings at a given point in time only account for the previous matches of a player, which is suboptimal in a retrospective analysis such as this where information on subsequent matches is available. As such, to estimate player skill levels, I apply the TrueSkill Through Time (TTT) algorithm of Dangauthier et al. (2007).

The TTT algorithm is based on the TrueSkill Bayesian model developed by Microsoft (Herbrich et al., 2006), which treats skill as a latent variable to be estimated and models the probability of win/loss outcomes based on the difference in player skills passed through a probit link. Specifically, suppose that player i and player j play a match at time t; at time t, their skills are denoted s_{it} and s_{jt} respectively. The TrueSkill model models the probability that player i beats player j as:

$$P\left(i \text{ beats } j | s_{it}, s_{jt}\right) = \Phi\left(\frac{s_{it} - s_{jt}}{\sqrt{2}}\right)$$

and each player's skill is assumed to follow a random walk with an initial condition:

$$s_{it_i^0} \sim \mathcal{N}\left(0, \sigma_0^2\right)$$
$$s_{i(t+1)} | s_{it} \sim \mathcal{N}\left(s_{it}, \gamma^2\right)$$

where t_i^0 is the time at which player i plays their first game.

The original TrueSkill algorithm calculates skill estimates at each time period using variational forward filtering, calculating the approximate posterior distribution of s_{it} conditional on all previous matches. The TTT algorithm extends this algorithm to perform forward filtering and backward smoothing iteratively until convergence, calculating the approximate posterior of s_{it} conditional on all previous and subsequent matches. This uses uses the information from all past and future games to inform the skill estimate of a player at any time point. I estimate player skills with the TTT algorithm, using all available data, with hours as the time unit (the most granular time unit at which data is available) using the hyperparameters $\sigma_0 = 1.02$ and $\gamma = 0.022$ (selected via grid search to maximize model evidence). Further details about the choice of hyperparameters and the results of the algorithm are given in Appendix B.3.

From this model, I obtain a posterior mean estimate of s_{it} for every player-hour pair that appears in the data, which I use in downstream analyses. I use these estimates in two ways. First, when analyzing player behavior (both in model free descriptives and in the behavioral analysis section), I consider skill to be a dimension of player heterogeneity and segment results by skill. Second, when I estimate the payoff function of the game, I control for player skill in the model, such that the estimated probability of one action winning over another is residual of the skills of the two players employing those actions. I note that, since the TTT model is estimated using only the win/loss outcomes of matches without accounting for action selection, the skill estimates will implicitly capture both in-game skill and skill at action selection; although these two dimensions of skill are likely to be highly correlated, they are not equivalent—some players may be very good at playing matches but bad at constructing teams, and vice-versa. This composite measure of skill will be correlated with the strategic value of an action (if a player is skilled at action selection, by definition they are more likely to have observations of actions with high win probabilities). However, the probability of a given action winning will still vary from game to game for a given player since they will sometimes match with players using actions with which they are relatively well- or poorly-matched. This allows for identification of the pairwise win probabilities between actions residual of skill (even though the

measure of skill includes partial information about action win probability).

4 Model-Free Descriptive Statistics

In this section, I present model-free descriptive statistics on player behavior. The goals of this section are twofold. First, I characterize the variability in the overall distribution of actions being used by players over time and assess the extent to which the game appears to converge towards a stable equilibrium or not. Second, I present stylized summaries of the individual-level variation in action selection relative to the population distribution, correlation to performance, and when and how players change their actions. These descriptive statistics show the macro dynamics of the game and give a suggestive picture of how players approach the decision problem. I segment many of the statistics by tercile of player skill to illustrate how these patterns vary directionally depending on how normatively good a player is.

4.1 Aggregate Dynamics of Action Distribution

I begin by quantifying the entropy of the population distribution of actions over time (Shannon, 1948). That is, denoting the theoretical distribution of action $\mathcal{X} = \{x_1, x_2, x_3, x_4, x_5, x_6\}$ on day t as $P_t(\mathcal{X})$, the entropy is:

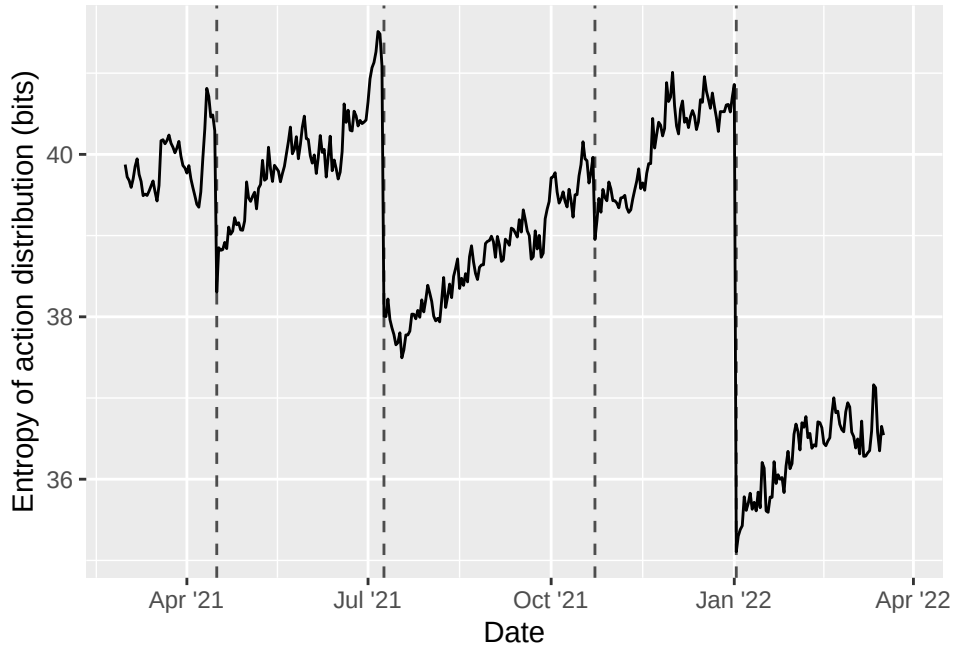
$$H_t = -\sum_{\mathcal{X}} P_t(\mathcal{X}) \log_2 P_t(\mathcal{X})$$

This gives a quantitative measure of how variable the distribution of actions being used in the game is at a given point in time. Mathematically, it gives a measure of how many bits of information are required, on average, to describe an action using an optimal compression code (optimal for the distribution $P_t(\mathcal{X})$). Intuitively, when the entropy is small, it says that usage is concentrated around a handful of actions that dominate gameplay; when entropy is large, it says that usage is spread across a large number of actions. However, given the sparsity of observations of $\mathcal{X}$ (i.e., most actions are only observed a few times in the entire dataset, let alone on a single day), the empirical entropy of the distribution over the exact action (set of six characters) is likely to be unreliable due to low statistical precision in empirical frequency estimates $\hat{P}_t(\mathcal{X})$. So, for these descriptive statistics, I approximate the entropy by treating each of the six characters within a team as an independent and identically distributed draw from a categorical distribution over the 702 characters in the data. That is, I approximate the distribution as:

$$P_t(\mathcal{X}) \approx \prod_{x_c \in \mathcal{X}} P_t(x_c)$$

where $P_t(x_c)$ is the probability that a random character drawn from a random team on day t is equal to x_c. Since this is a categorical distribution over a relatively low-dimensional object, it can be accurately approximated by its empirical distribution. Since entropy is additive across independent

Figure 3: Entropy of action distribution over time



Note: the vertical lines indicate dates of rule changes. The first rule change shrank the action space (going from one to zero restricted characters allowed and banning Dynamax), the second and fourth changes grew the action space (going from zero to one then one to two restricted characters allowed), while the third change kept the action space constant (at one restricted character allowed) while changing the payoff function by permitting Dynamax.

events, this yields the estimated entropy:

$$\hat{H}_t = -\sum_{\mathcal{X}} \sum_{x \in \mathcal{X}} \hat{P}_t(x) \log_2 \hat{P}_t(x)$$

Figure 3 shows the time series of action distribution entropy calculated at the daily level. The entropy of the distribution tends to drop substantially immediately after each rule change, then gradually increases until the next rule change. This result is rather surprising. One might expect that entropy is highest immediately following a rule change, with players being unsure which actions to use, followed by a decrease as players identify the best actions and converge towards an equilibrium. The entropy is also quite large in absolute scale (ranging from 35 to 41.5 bits), indicating that players need to account for a large number of possible opposing actions if they want to maximize their probability of winning.

The initial drop in entropy followed by a gradual increase may suggest that players at first choose "safe" actions—such as actions more familiar to the player or more robust to changes in the overall action distribution—after a rule change, and then gradually explore more diverse and tailored actions once they are familiar with the current distribution under the new rules.

The aggregate patterns are, in part, driven by heterogeneity in action usage by player skill: the

left panel of Figure 4 shows that skilled players concentrate around a smaller set of actions, with the top tercile of players having much lower action distribution entropy than the bottom tercile. The right panel shows that skilled players play most heavily immediately after a rule change, explaining part of the drop in overall entropy at each rule change. Entropy being lower for skilled players may indicate that skilled players are less prone to making noisy or idiosyncratic decisions, leading to the distribution concentrating on a smaller subset of normatively well-adapted actions. The fact that skilled players disproportionately play the most at the beginning of each ruleset may also indicate that skilled players get bored of the game and drop out faster without the periodic rule changes to keep them engaged.

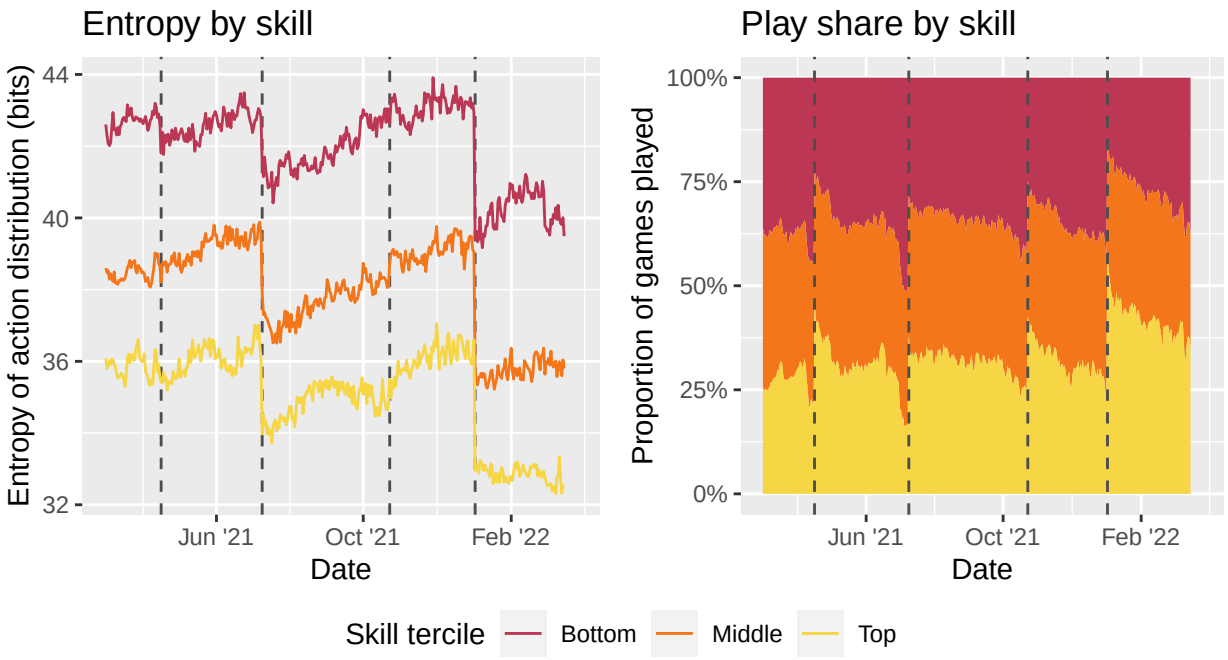
Even within skill level, there is substantial non-stationarity in entropy. In particular, in response to the second and fourth rule changes, where limits on the number of allowable "restricted" characters were relaxed, entropy sharply dropped, and more steeply for more skilled players. Conversely, after the first rule change, when fewer restricted characters were allowed, low-skill players' entropy dropped slightly, while it stayed about constant for high-skill players. The increasing entropy over time within a ruleset is also apparent within skill terciles, though it is less prominent than the aggregate trend. Notably, after the third rule change, where the set of allowable characters was not changed, but rather a mechanic that can make any character more powerful (Dynamax) was allowed, entropy actually increased slightly.

These results show that it is not only the sheer number of actions available in the game but also their degree of differentiation that influences how diffuse the action distribution is. While allowing restricted characters increases the absolute number of possible actions legal in the game, restricted characters are usually much stronger than non-restricted characters, increasing vertical differentiation between actions, in that actions that include the maximum number of restricted characters are generally better than actions that do not. This differentiation presumably makes it easier for players to identify relatively good actions, resulting in lower entropy (especially for skilled players). Conversely, Dynamax can make any character stronger, which in principle should decrease differentiation between actions. Thus, players choose from a larger number of actions that yield similar performance, increasing entropy.

Even during periods where the overall entropy of the action distribution is relatively stable, the actual distribution of character usage is often still evolving. This is apparent when looking at the usage rates of individual characters. Figure 5 shows the usage rates for two illustrative characters. Even weeks after a rule change, the usage of characters can change substantially, sometimes quite suddenly. Additionally, while some characters go through the most significant usage changes immediately after a rule change, others appear to have a somewhat more delayed reaction.

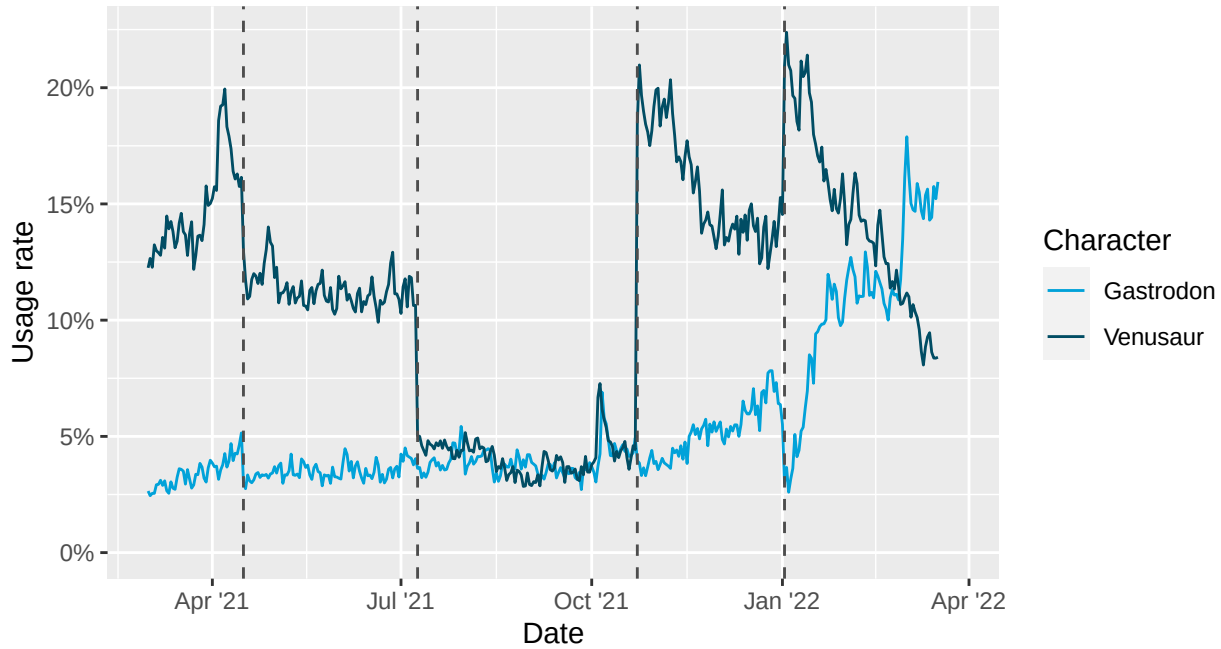
Taken together, these results show that the decision problem faced by players is complex and evolving. To select an action that will perform well overall in the game, a player needs to consider their chance of winning marginalized over the highly entropic distribution of opposing actions. This distribution is evolving over time, compounding the difficulty of the problem. Rather than starting from a diffuse set of naive actions and then gradually converging towards a more concentrated

Figure 4: Action distribution entropy, segmented by skill



Note: vertical lines indicate dates of rule changes. Entropy figures are calculated using the empirical distribution within each skill tier on each day.

Figure 5: Usage patterns of illustrative characters



Note: the vertical lines indicate dates of rule changes. "Usage rate" is the proportion of actions on a given day that include the character in question. Neither Gastrodon nor Venusaur are considered restricted characters.

equilibrium, it seems that players start from a smaller set of "safe" actions, then become more exploratory as they become familiar with a ruleset. These directional patterns hold across players of all skill levels, though skilled players concentrate around a smaller set of actions, possibly indicating that they are closer to an equilibrium. Skilled players also appear to become bored of a ruleset and reduce play volume faster than less skilled players.

4.2 Individual-Level Dynamics of Actions and Performance

To further investigate the individual-level dynamics of these aggregate patterns, I calculate the Shannon information of each action (negative log-likelihood, in units of bits) based on the empirical action distribution for that day. As in the entropy calculations, I approximate the six characters within a team as being independent and identically distributed. I will hereafter refer to the Shannon information as the "atypicality" of an action.

Specifically, for a given action $\mathcal{X}$, I estimate its Shannon information/atypicality at time t as:

$$\hat{I}_t(\mathcal{X}) = -\sum_{x \in \mathcal{X}} \log_2 \hat{P}_t(x)$$

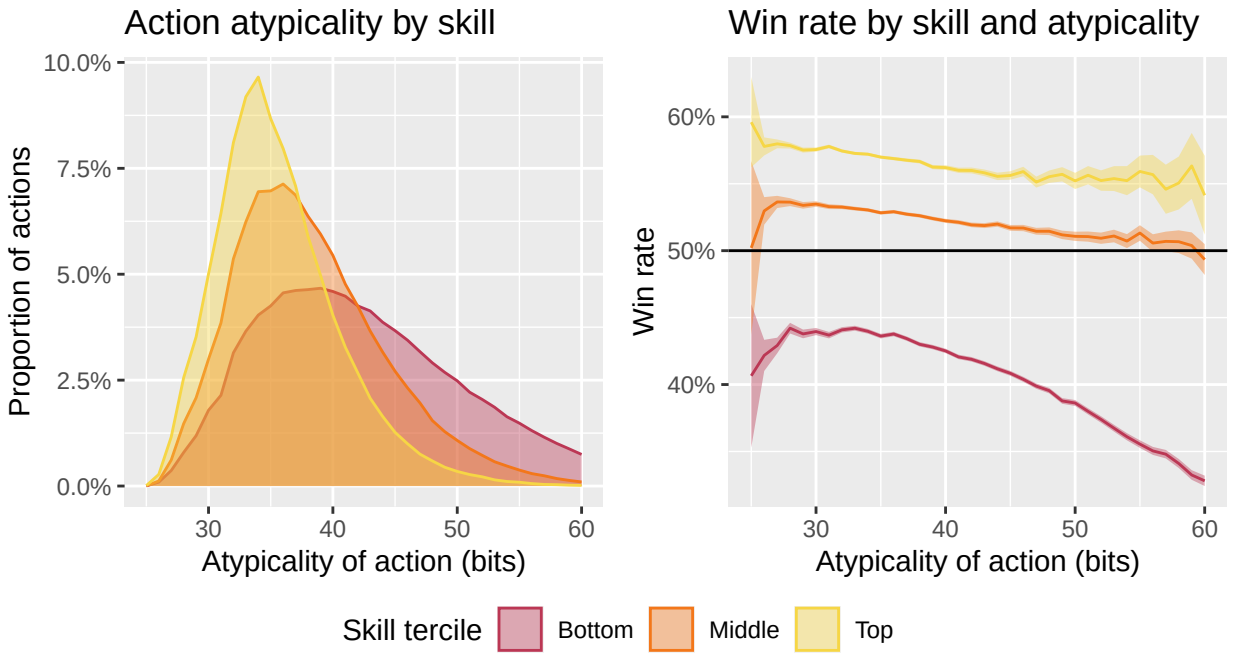
Mathematically, this says how many bits of information are needed to describe action $\mathcal{X}$ under an optimal compression for the distribution $P_t(\mathcal{X})$. Intuitively, smaller values indicate an action that is relatively likely to occur under the current distribution, while larger values indicate unlikely actions. Note that while atypical actions by definition have small probability individually, atypical actions may still collectively comprise substantial mass in the distribution of atypicality if there are many of them.

Figure 6 visualizes, by skill level, the distribution of action atypicality and the relationship between atypicality and empirical performance. The smooth distributions with interior modes show that there is smooth variation in how popular actions are, rather than, e.g., a bimodal distribution split between a handful of extremely popular actions and a large number of extremely niche actions. Higher-skill players tend to concentrate around more typical actions, with lower mean and variance in atypicality. Meanwhile, lower-skill players have a longer tail of more atypical actions.

The relationship between win rate, player skill, and action atypicality suggests that, within skill level, more typical teams tend to perform better, but there is also a large level shift in win rate between players of different skill levels. Skilled players tend to choose more typical actions, which tend to normatively perform better; and, conditional on the typicality of their action, they tend to perform better with that action. Thus, it appears that skilled players are better able to identify and concentrate on normatively well-adapted actions, while lower-skill players choose their actions more naively or idiosyncratically. This is consistent with the discussion in Section 3.3 about my skill measure capturing two dimensions of skill simultaneously: skill at selecting a good action, and skill at performing well in-game conditional on the selected action.

Next, I look into the dynamics of individual-level adjustment of actions over time. Since I only observe timestamps at the hourly level, the sequence of matches for a given player is unknown within

Figure 6: Performance by player skill and action typicality



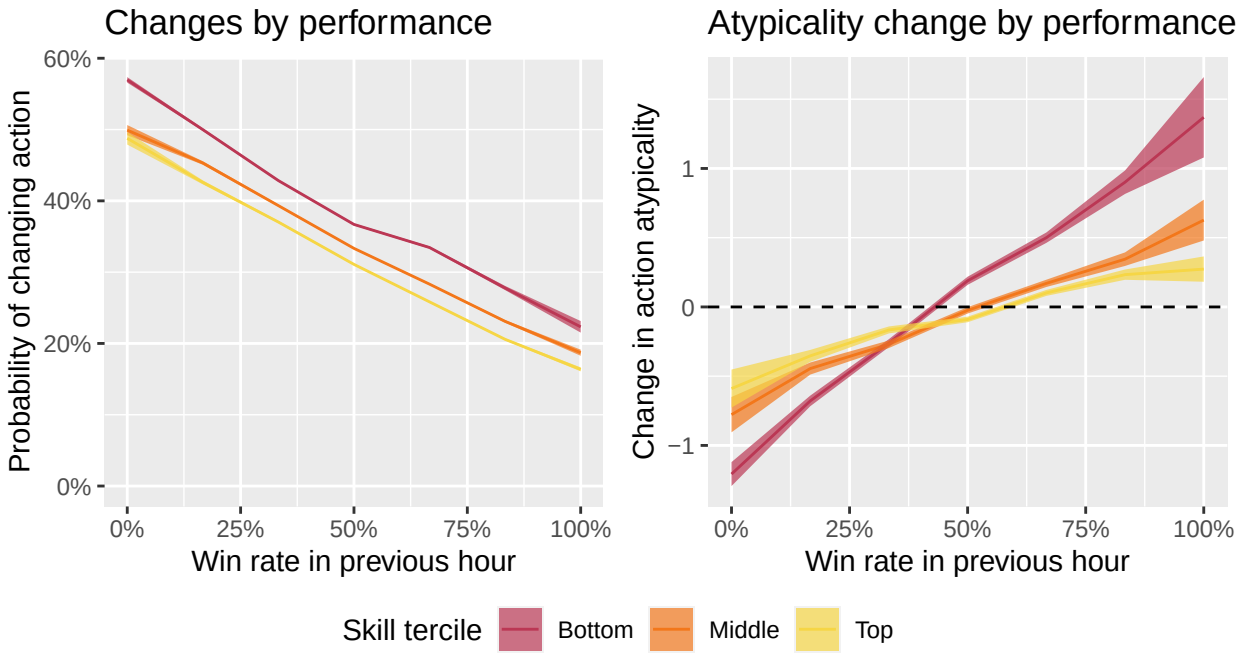
Note: the bands in the right panel represent pointwise 95% confidence intervals. Both plots are calculated by discretizing atypicality into bins of unit width.

a given hour, so I aggregate observations to the hourly level. Specifically, I calculate the modal action within each player × hour pair (the most frequently occurring assortment of six characters), clean the data to ensure that each hourly observation contains only match outcomes corresponding to a single action for that player, and then calculate the empirical win rate within that hour.[13] I define an action change as having occurred when at least one character is changed from the previous hour observation for that player. I remove the first hour that each player played in a given ruleset (since an action change is almost certain to occur between rulesets). This results in 5,239,825 player-hour observations across 219,190 players, with an average of 24 observations per player and a median of 4. Each hourly observation is, on average, the aggregation of 3.96 match outcomes.

Figure 7 shows, as a function of skill and recent performance, how likely players are to change their actions and, conditional on making a change, whether they tend to move closer to or further from the current action distribution (as captured by the change in action atypicality). The left panel shows strong evidence that players learn by reinforcement: players are more than twice as likely to change their action after a streak of losses as opposed to a streak of wins. This pattern is remarkably consistent across skill levels, with small level-shifts (higher-skilled players are less likely to change overall). On average, players change their actions in 34.02% of player-hour observations.

[13]Further details on this data cleaning and aggregation are given in Appendix B.2. In brief, if a non-modal action matches the modal action in the preceding/succeeding hour of play for a given player, I reassign the corresponding action and match outcome to that adjacent hour. This results in 3.5% of actions being reassigned to adjacent hours, such that 95.0% of actions overall correspond to the modal action for that hour. I remove the observations corresponding to the remaining 5.0% of actions (but only for analyses requiring hourly aggregation).

Figure 7: Player response to recent performance



Note: the bands represent pointwise 95% confidence intervals. Both panels subset to observations where there were at least five matches in the previous hour of play. Win rate is discretized into six bins with breaks from 1/12 to 11/12 spaced 1/6 apart.

Additionally, while Figure 6 showed that more typical actions tend to perform better, players adjust their actions to be more or less typical depending on their recent performance, as seen in the right panel. After a losing streak, players tend to change their actions to be more in line with other players, but after a winning streak, players tend to deviate away from popular actions.

Low-skill players are especially prone to this behavior, tending to deviate substantially away from the population distribution (i.e., changing their action to be much more atypical) after positive reinforcement, despite atypical actions tending to achieve worse performance. Skilled players exhibit the same directional pattern, but are less sensitive to recent performance (indicated by the smaller slope) and tend to lean more towards imitation than deviation (indicated by the lower intercept). This may indicate that low skill players partially perform worse due to overreaction to their recent performance, easily becoming overconfident or overly pessimistic about their own ability to effectively select actions that deviate from what is typical. This is consistent with prior findings that, for instance, professional investors achieve better investment outcomes than lay investors due partially to better management of their affective response to gains and losses (Chu et al., 2014).

4.3 Discussion of Descriptive Findings

While the above descriptive patterns are suggestive, it is difficult to draw conclusions about the decision rules driving player behavior or assess the normative properties of their actions based on raw data alone. For instance, while the left panel of Figure 7 suggests that players account for

performance when deciding whether to change actions, it does not disentangle the different ways players may account for performance information: do they react only to directly observed win/loss outcomes, or are they able to reason more broadly about the normative fitness of their action beyond the specific opponents they recently matched against? Are they able to correct for factors outside their control such as their opponents' skill and stochasticity in the outcome of the game? Likewise, in terms of action selection, while the right panel of Figure 7 shows that players tend to imitate popular actions after losing and deviate towards less popular actions after winning, this could be for several reasons. For instance, players may imitate overall popular actions, or they may specifically imitate the opponents they lost to, or they may reason strategically about how to best respond to their opponents (with the best response tending to be more in line with the current action distribution). In terms of normative implications, though Figure 6 shows that more atypical teams tend to perform worse, this combined with Figure 7 does not necessarily mean that players are behaving suboptimally by selecting less popular actions after a winning streak, since there are many characteristics of an action besides its overall typicality which may affect its normative fitness.

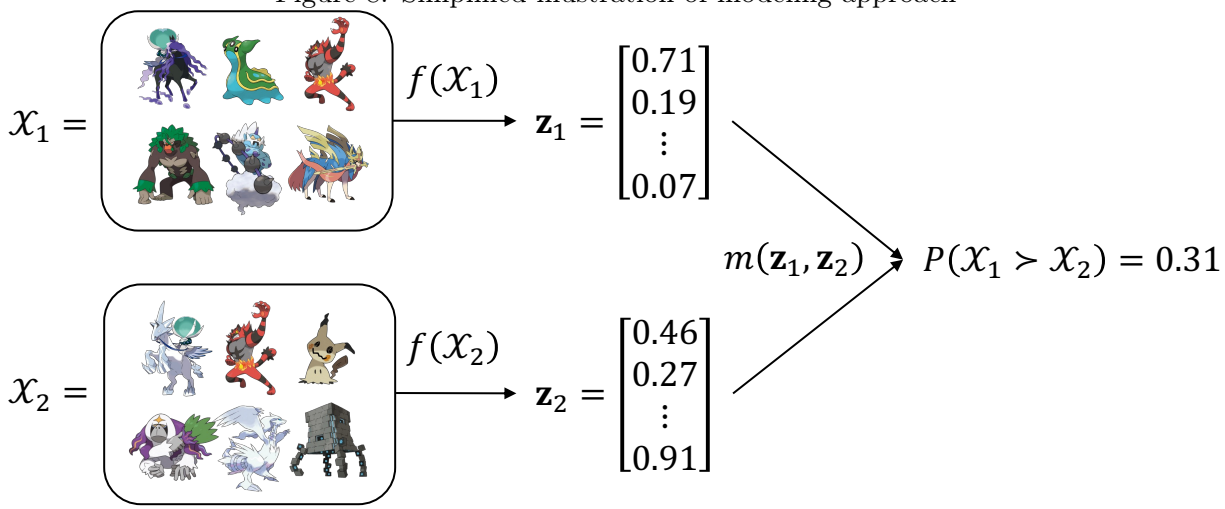
Thus, to be able to draw conclusions about the behavioral properties of player decisions, it is necessary to incorporate different factors that may be driving those decisions jointly into a model. For instance, for analyzing how players decide whether to change actions, the binary decision can be modeled in a regression as a function of variables such as the empirical win rate, overall normative fitness (overall win probability of an action, marginal of opposing action), skill difference relative to opponents, etc. For analyzing how players select actions conditional on making a change, the chosen action can be modeled as a function of variables such as the actions of recent opponents, the overall distribution of actions, the actions that best respond to each of these, etc. The challenges with implementing such models are twofold. First, quantities like normative fitness and best responses require knowing the pairwise win probabilities between pairs of actions; second, due to the complexity of the action space, it is unclear how best to specify a regression model of player choices over possible actions. In the next section, I propose a machine learning model which addresses both challenges simultaneously.

5 Supervised Representation Learning Model of Win Probabilities

5.1 Task Structure and Model Desiderata

As discussed above, while the descriptive results in Section 4 are suggestive, it is difficult to draw normative and behavioral conclusions based on the raw data alone. The payoff structure of the game (i.e., probability that one action wins over another, after controlling for skill) is not known a priori and so must be empirically estimated. However, the action space is combinatorially large and the payoff structure is complex. Thus, it is necessary to model the payoff structure with a method that scales to high-dimensional inputs and is sufficiently expressive to capture complex interactions between inputs. At the same time, for downstream behavioral analysis, it is necessary to map the high-dimensional discrete action space to a more mathematically tractable representation that is

Figure 8: Simplified illustration of modeling approach



Player actions $\mathcal{X}_1$, $\mathcal{X}_2$ (discrete sets with a combinatorially large number of possible levels) are passed through embedding function $f(\cdot)$ to obtain representations $\mathbf{z}_1$, $\mathbf{z}_2$ (numeric vectors), which are then combined via matchup function $m(\cdot, \cdot)$ to output a win probability. Note that this illustration omits the skill correction term for simplicity.

conducive to downstream modeling. I propose a supervised representation learning approach that addresses both issues.

First, I formalize the modeling task and lay out the basic structure of my proposed model, with description of the exact model specification deferred to Section 5.2. Specifically, suppose that Player 1 and Player 2 select actions $\mathcal{X}_1$ and $\mathcal{X}_2$, and that their skill levels are s_1 and s_2 respectively (as estimated by the TTT model discussed in Section 3.3). I first map the actions $\mathcal{X}_1$, $\mathcal{X}_2$ (which are discrete variables with a combinatorially large number of possible values) to numeric vector values (embeddings) $\mathbf{z}_1 = f(\mathcal{X}_1)$, $\mathbf{z}_2 = f(\mathcal{X}_2)$ using an *embedding function* $f(\cdot)$; then, I map the embeddings $\mathbf{z}_1$, $\mathbf{z}_2$ to Player 1's log-odds of winning using a *matchup function* $m(\cdot, \cdot)$ paired with an additive skill correction. Specifically, I model the probability that Player 1 wins as:

$$P(\mathcal{X}_1 \succ \mathcal{X}_2 | s_1, s_2) = \sigma\left(m\left(f(\mathcal{X}_1), f(\mathcal{X}_2)\right) + w_s(s_1 - s_2)\right)$$

where $\sigma(\cdot) = \frac{\exp(\cdot)}{1+\exp(\cdot)}$ is the sigmoid or inverse logit function. An intuitive illustration of this basic model structure is given in Figure 8. In Appendix C.2, I consider an extension of the model where player skill enters non-additively via the f function, but find that it does little to improve the explanatory power of the model.

A key modeling decision is how to parameterize the embedding function $f(\cdot)$ and the matchup function $m(\cdot, \cdot)$. Without further restrictions, the two functions are not separately identified: for instance, for any invertible function $h : \mathcal{X} \mapsto \mathbf{z}$, the alternative parameterization $\tilde{f}(\cdot) = h(f(\cdot))$ and $\tilde{m}(\cdot, \cdot) = m\left(h^{-1}(\cdot), h^{-1}(\cdot)\right)$ will yield identical predictions to the original f and m but imply different intermediate embeddings $\tilde{\mathbf{z}} = \tilde{f}(\mathcal{X})$; the embeddings from different observationally equivalent models may be more or less amenable to behavioral interpretation.

My approach to resolve this identification issue is to restrict the function m to enforce certain

desiderata on the properties of the embeddings while leaving f as a flexible nonparametric function; this then compels the embedding function f to map the action space to embeddings that satisfy the desiderata. Specifically, for interpretability, I want the matchup function m to have the following properties with respect to the embedding space:

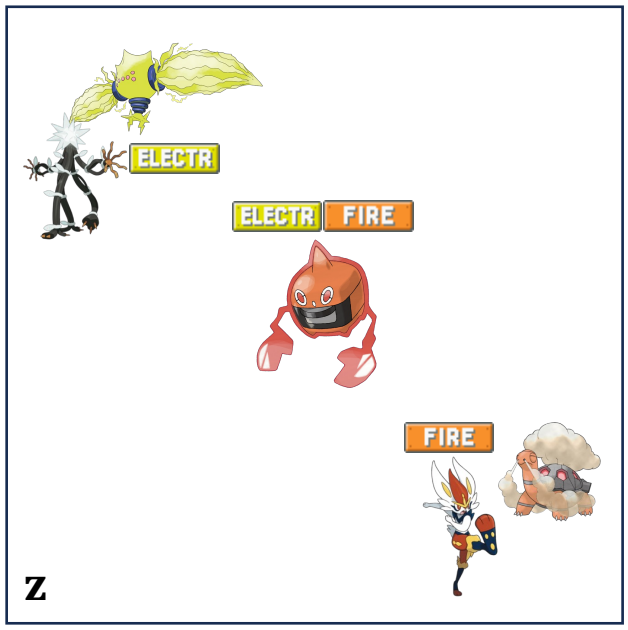
1. **Smoothness and uniqueness of output w.r.t. embeddings:** for two given input embeddings $\mathbf{z}_1$, $\tilde{\mathbf{z}}_1$, it should be the case that $\mathbf{z}_1 \approx \tilde{\mathbf{z}}_1$ if and only if $m(\mathbf{z}_1, \mathbf{z}_2) \approx m(\tilde{\mathbf{z}}_1, \mathbf{z}_2)$ for all opponent embeddings $\mathbf{z}_2$. In plain words, this means that if two actions $\mathcal{X}_1$, $\tilde{\mathcal{X}}_1$ have similar strategic properties (i.e., they tend to have similar probabilities of winning against any given opponent), they should be close together in the embedding space. Conversely, if two actions are close together in the embedding space, they should have similar strategic properties. This ensures that measures of distance and similarity in the embedding space (e.g., Euclidean distance and cosine similarity) capture the similarity between actions in terms of their strategic properties.

2. **Meaningful linear combinations of embeddings:** for two given input embeddings $\mathbf{z}_1$, $\tilde{\mathbf{z}}_1$, it should be the case that $m(\alpha\mathbf{z}_1 + \tilde{\alpha}\tilde{\mathbf{z}}_1, \mathbf{z}_2) = \alpha \cdot m(\mathbf{z}_1, \mathbf{z}_2) + \tilde{\alpha} \cdot m(\tilde{\mathbf{z}}_1, \mathbf{z}_2)$ for all opponent embeddings $\mathbf{z}_2$ and scalar coefficients α, $\tilde{\alpha}$. In plain words, this means that if an action has an embedding that linearly combines the embeddings of two other actions, the action should likewise linearly combine the strategic properties of the two other actions (i.e., its probability of winning or losing against a given opponent will combine the win/loss probability of the two actions it is combining). This ensures that interpolations between different actions in the embedding space will correspond to meaningful interpolations between their strategic properties.

These desiderata, which are illustrated in a stylized example in Figure 9 for intuition, are essential for behavioral analysis. In practice, players will tend to make noisy or imprecise decisions, and so may choose an action that approximately (but imperfectly) corresponds to a decision rule. The first desideratum ensures that, for instance, if a player chooses an action that has similar (but not exactly the same) strategic properties as another action implied by a given decision rule, the embedding of the chosen action will be close to the embedding of the action implied by the decision rule. This way, in downstream behavioral analyses, the player's imprecision can be modeled as an error term in the embedding space, with the variance of the error term capturing how imprecise their decision-making process is.

Additionally, players may employ a combination of decision rules: for instance, players may sometimes imitate the actions of their past opponents and sometimes calculate the best response to their opponents, or they may select an action which combines some strategic properties of imitation and best response. The second desideratum ensures that actions which interpolate between the strategic properties of different actions are also mapped to interpolating points in the embedding space. This ensures that, in downstream behavioral analyses, combinations of decision rules can be modeled as linear combinations in the embedding space.

Figure 9: Simplified illustration of model desiderata



Characters with similar strategic properties (e.g., electric types or fire types) should be clustered together in the embedding space (smoothness) and characters that combine the properties of these clusters (e.g., dual electric/fire types) should fall in between the clusters in the embedding space (linear interpolation). Note that this is a highly stylized illustration, since in reality, an action consists of a set of six characters rather than a single character, and there are numerous other character attributes besides type that determine their strategic properties.

Combining these two desiderata ensures that, for instance, a linear regression modeling a player's action selection as a linear combination of the actions implied by different decision rules in the embedding space can be meaningfully interpreted: the coefficients can be interpreted as the weight players put on different decision rules, while the error variance can be interpreted as how imprecise players are at implementing these decision rules.

With the basic task structure and desiderata laid out, I next discuss my proposed model for enforcing these desiderata.

5.2 Representation Learning Model Specification

To enforce the above desiderata, I specify the matchup function as a bilinear form $m(\mathbf{z}_1, \mathbf{z}_2) = \mathbf{z}_1' \mathbf{M} \mathbf{z}_2$ for some weight matrix $\mathbf{M}$. Intuitively, the dimensions of embedding vectors $\mathbf{z}_1$, $\mathbf{z}_2$ can be thought of as different latent attributes of the actions $\mathcal{X}_1$, $\mathcal{X}_2$ while the weight matrix $\mathbf{M}$ can be thought of as capturing the strategic matchups between each pair of attributes. For instance, if $\mathbf{M}_{ij}$ is a large positive value, then this means that, all else equal, Player 1 will be more likely to win if their action embedding $\mathbf{z}_1$ loads higher on the i-th dimension and the opponent's action embedding $\mathbf{z}_2$ loads higher on the j-th dimension (e.g., in my context, dimension i may capture how heavily an action uses water-type attacks while dimension j captures how heavily an action uses fire-type characters). Likewise, a negative value means that there is a strategic disadvantage to dimension i if the opponent loads highly on dimension j. Note that this functional form does not assume that

the logit probability of winning is linear in the original actions $\mathcal{X}_1$ or $\mathcal{X}_2$; rather, it forces the model to learn a nonlinear transformation $\mathbf{z} = f(\mathcal{X})$ such that the resulting transformed variable admits a linear structure.

This functional form, by construction, satisfies the desideratum that linear combinations be meaningful in the embedding space, since m will be linear in each argument conditional on the other argument. It further satisfies the desideratum that the output be smooth and unique with respect to the embeddings so long as the matrix $\mathbf{M}$ is invertible. In particular, in Appendix C.1, I show that under the bilinear functional form, as long as $|m(\mathbf{z}_1, \mathbf{z}_2) - m(\tilde{\mathbf{z}}_1, \mathbf{z}_2)| \leq \varepsilon \ \forall \mathbf{z}_2 \in \Delta^K$, then $\max_{k=1,\dots,K} |\mathbf{z}_{1k} - \tilde{\mathbf{z}}_{1k}| \leq K \max_k \sum_j \left|\mathbf{M}^{-1}_{jk}\right| \varepsilon$, where $\left|\mathbf{M}^{-1}_{jk}\right|$ is the absolute value of the j,k-th element of $\mathbf{M}^{-1}$. The condition that two actions have similar log-odds of winning against all opposing opponents in a unit simplex in the embedding space is without loss of generality: as long as the ε bound holds over any K linearly independent opposing $\mathbf{z}_2$, an analogous bound on $\max_{k=1,\dots,K} |\mathbf{z}_{1k} - \tilde{\mathbf{z}}_{1k}|$ holds. In other words, this bound shows that as long as two actions $\mathcal{X}_1$, $\tilde{\mathcal{X}}_1$ are strategically similar (in that their log-odds of winning against a given opponent are no more than ε different for many possible opponents $\mathcal{X}_2$), then their embeddings must also be close in that no coordinate will differ by more than a quantity proportional to ε.

This shows that, if a pairing of an embedding function $\mathbf{z} = f(\mathcal{X})$ and matchup function $m(\mathbf{z}_1, \mathbf{z}_2) = \mathbf{z}_1' \mathbf{M} \mathbf{z}_2$ exist such that they jointly reproduce the log-odds of winning between all pairs of actions, then the embeddings will satisfy the above desiderata. It is not necessarily guaranteed that such a pairing exists; however, I show below in Section 5.3 that parameterizing both f and m as generic feed-forward neural networks (which are known to approximate any continuous function up to arbitrary accuracy with sufficient network width) does not improve empirical model performance, demonstrating that the assumption of bilinearity is not too restrictive.

This functional form restriction plays an analogous role to the dot product used in embedding-based language models for predicting the co-occurrence of words, such as word2vec and transformer-based models (Mikolov et al., 2013; Vaswani et al., 2017). There, each token (unique word) is represented by a "word embedding" vector, a target word's context (e.g., the words adjacent to it in a sentence) is reduced to a "context embedding" vector, and a given token's probability of occurring in a given context is increasing in the dot product between these two vectors. The bilinearity of the dot product ensures that the embedding space has a linear structure wherein linearly combining words will combine the contexts in which those words occur: for instance, in a well-known example from Mikolov et al. (2013), the word2vec model learns embeddings such that $\mathbf{z}_{queen} \approx \mathbf{z}_{king} + \mathbf{z}_{woman} - \mathbf{z}_{man}$, showing that the implied gendering of a word can be changed through simple linear combinations in the embedding space between words with different genderings. This is because the dot product ensures that such linear combinations are meaningful, since the bilinearity of the dot product means that this linear combination will correspond to a word which appears in contexts that combine the contexts where "king" and "woman" appear while removing the context where "man" appears. This bilinear structure ensures that linear combinations in the embedding space are semantically meaningful, such that linear combinations and distance/similarity measures are

interpretable.

In my model, the bilinear matchup function likewise ensures that linear combinations in the embedding space are strategically meaningful, since they will correspond to linear combinations between which opponent actions they will tend to perform well/poorly against. Thus, in the same way that word embeddings allow for intepretable linear combinations and distance/similarity measures between words in terms of semantic meaning, my proposed use of a bilinear functional form allows for interpretable linear combinations and distance/similarity measures between actions in a game in terms of strategic properties.

My full model specification is as follows: I parameterize f as a fully connected feed-forward neural network with one 100-dimensional hidden layer with ReLU activations and a 100-dimensional output layer with sigmoid activations (thus yielding embeddings in the unit hypercube). The inputs for a given action include six categorical variables each with 703 categories (for each unique character in the data)[14] and a binary indicator for whether the "Dynamax" mechanic is allowed under the current ruleset (since the strategic attributes of a character may depend on whether it is able to use this mechanic or not). I then parameterize m as a bilinear form over the two 100-dimensional outputs of the f network for the two action inputs.

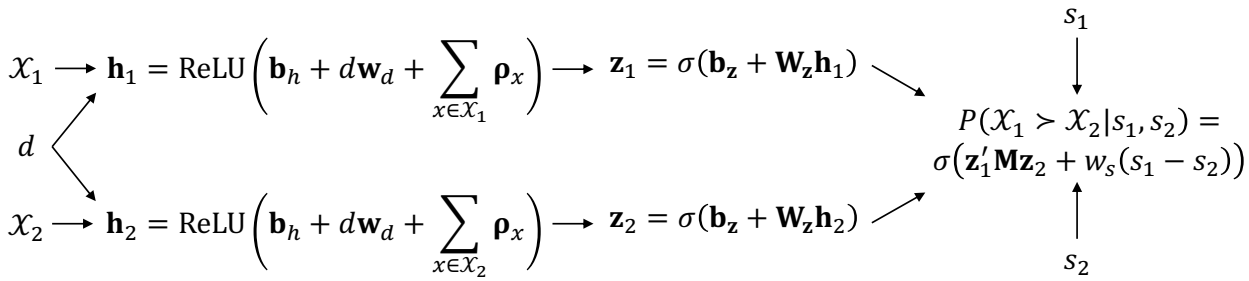
For parsimony, I also introduce constraints on the model parameters that ensure invariance of the model outputs to the (arbitrary) ordering of players and characters (Zaheer et al., 2017). Specifically, since the six characters within a team are exchangeable, I constrain the weights in the first layer of the neural network to be symmetric for the six categorical variables. Likewise, since the game is symmetric, the index of which player is "player 1" or "player 2" is arbitrary, and so I constrain the weight matrix $\mathbf{M}$ to be antisymmetric (such that $\mathbf{M} = -\mathbf{M}'$). This ensures that the model returns the probability complement when the positions of the two players are flipped. The full architecture is summarized in Figure 10 and contains 85,551 free parameters.

Note that, while the specific inputs and constraint structures of the model are somewhat idiosyncratic to my empirical context, this basic model specification can easily be extended to other types of games. While in my empirical context the game has symmetric action spaces and payoff structure, asymmetric games (where the action space and/or payoff structures are different for the two players) could be similarly modeled by fitting separate embedding networks f_1, f_2 that map the two players' actions to the same embedding space and removing the constraint that the weight matrix $\mathbf{M}$ is antisymmetric. This could apply, for instance, to settings such as team drafting in sports, where teams have differing budgets for player contracts and differing orders of draft picks, resulting in a different sets of feasible actions (and since the utility of a player depends on that player's fit with an existing team, the payoff will also be asymmetric).

Similarly, while the payoff is zero-sum in my case (such that the expected outcome can be represented by a single-dimensional output), games where the players have separate payoffs may be

[14] Though actions generally consist of six unique characters, it is technically legal to select an action which contains only four or five characters, though there is generally no strategic benefit to doing so; in such cases, I treat an "empty" character slot as an additional category, resulting in 703 total categories. Less than 1.2% of actions contain fewer than six characters.

Figure 10: Visual diagram of proposed neural network architecture



Note: $\mathcal{X}_1$, $\mathcal{X}_2$ represent the discrete set of six characters of the two players, while s_1 and s_2 are scalar estimates of their skill at the time of the match as calculated by the TTT model. d is a binary indicator of whether the dynamax mechanic is allowed under the ruleset in effect at the time of the match. The activation functions are defined element-wise as $\text{ReLU}(\mathbf{x})_k = \max(0, \mathbf{x}_k)$ and $\sigma(\mathbf{x})_k = \exp(\mathbf{x}_k) / (1 + \exp(\mathbf{x}_k))$. The matrix M is constrained to be antisymmetric, while all other parameters are left as fully free parameters (except for L_2 regularization). In my preferred specification, hidden layers$\mathbf{h}_i$ and embeddings $\mathbf{z}_i$ are 100-dimensional.

represented by stacking multiple bilinear forms (e.g., player 1's payoff is modeled as $\mathbf{z}_1'\mathbf{M}_1\mathbf{z}_2$ while player 2's payoff is modeled as $\mathbf{z}_1'\mathbf{M}_2\mathbf{z}_2$). This is applicable, for instance, in competitions that have a coordination element, such as two firms developing products where the total demand is higher when the products are complementary with each other.

Likewise, while I analyze a two-player game, generally, n-player games where there are more than two players may be modeled to satisfy the same desiderata by employing a multilinear parameterization, where $m(\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n)$ is modeled as an n-dimensional tensor product. Thus, the approach of using a bilinear (or multilinear) parameterization of the game outcomes as a function of action embeddings may be applied to any game with a large action space and/or complex payoff structure, enabling behavioral analysis of player actions in the embedding space.

5.3 Model Training and Empirical Performance

I train the f and m models jointly to maximize the log-likelihood of the observed win outcomes. I randomly split the matches in the data into 80% training, 10% validation, and 10% test sets. I perform stochastic optimization using the Adam stochastic gradient descent algorithm with minibatches of 1,000 using a learning rate of 0.002, L_2 regularization factor of 3×10^{-6}, and default momentum parameters of 0.9 and 0.999 (Kingma and Ba, 2014). I train for up to 150 epochs, retaining the parameter estimates from the epoch that achieved the best validation likelihood. I selected the learning rate, regularization factor, and activation functions that resulted in the best test likelihood. Adding another hidden layer to the embedding function did not improve performance. See Appendix C.2 for results of other hyperparameters and model specifications.

Some summaries of the model performance are given in Table 2. Overall, the model obtains an out-of-sample McFadden pseudo-R^2 of 12.75% (McFadden, 1974); while most predictive performance is explained by a simple logistic regression of win outcome on skill difference (11.64%), the action information still captures substantial additional variation. Compared to a model that parameterizes

Table 2: Performance of proposed model and alternatives

Model	Data	Parameters	NLL	McFadden R^2	Accuracy	AUC
Proposed bilinear	Train	85,551	0.6015	13.22%	66.75%	73.37%
	Validation		0.6037	12.91%	66.57%	73.10%
	Test		0.6048	12.75%	66.43%	72.96%
Standard feed-forward	Train	90,701	0.6017	13.19%	66.73%	73.35%
	Validation		0.6038	12.89%	66.58%	73.09%
	Test		0.6049	12.73%	66.40%	72.94%
Skill-only logistic	Train	1	0.6119	11.72%	65.64%	71.97%
	Validation		0.6113	11.81%	65.75%	72.07%
	Test		0.6125	11.64%	65.54%	71.89%

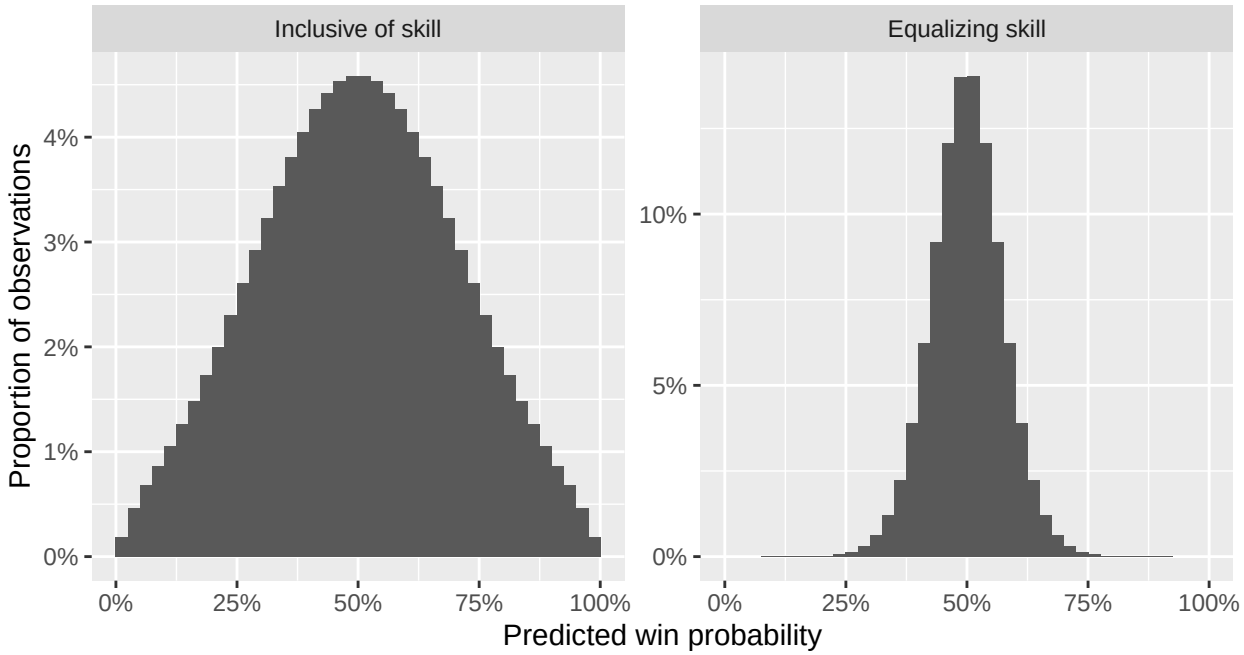
Note: NLL refers to per-observation negative log-likelihood, accuracy refers to the proportion of the time the winning player was predicted to have >50% probability of winning by the model, and AUC refers to the area under the receiver operating characters (ROC) curve. "Standard" refers to a more conventional feed-forward neural network specification (described in Appendix C.2). "Skill-only" refers to a simple logistic regression (with no intercept) on the difference in TTT skill estimates (as described in Section 3.3) between the two players.

$m(\cdot, \cdot)$ as a feed-forward neural network with one hidden layer of 100 units (see Appendix C.2 for full specification details), the bilinear specification achieves nearly identical (even slightly better) performance. Thus, assuming a bilinear form on the final layer of the neural network does not worsen performance, while also having the added benefit of yielding embeddings with an interpretable linear structure that can be used for downstream analysis.

Figure 11 shows the distribution of predicted win probabilities from the model, before and after controlling for skill (i.e., the right panel shows the hypothetical probability of one action winning over another if the players using those actions had equal skill). Though skill captures most of the variation in win probabilities, there is still meaningful variation in win probabilities even after controlling for skill. The win probability distribution has a standard deviation of 20.4% before equalizing skill and 7.41% after, with the advantaged player having a 66.6% chance of winning on average before equalizing skill and 55.8% after equalizing skill. The win probabilities before and after equalizing skill have a sample correlation of 41.3%. Overall, these results show that although in-game skill and quality of selected action are correlated, skill does not fully capture variation in win probabilities, and the specific actions chosen by the player are still consequential.

Note that it is unsurprising that the skill-only model performs so well: the TTT model estimates time-varying, player-level parameters specifically optimized to predict win outcomes. However, the skill model by itself does not allow for behavioral analysis of actions as it does not capture the pairwise interactions between the actions selected by the two players. Though the improvement in predictive performance from adding action information is relatively small, the goal is not to maximize predictive accuracy of the model but rather to estimate representations that meaningfully capture the pairwise strategic interactions between actions above and beyond what can be explained by skill. Figure 11 shows that there is still substantial additional variation in pairwise win probabilities after controlling for skill, and it is this additional variation that allows for the learning of action representations and for subsequent behavioral analysis.

Figure 11: Distribution of model predicted values



Note: the left panel shows the distribution of model output (win probability) $\hat{P}(\mathcal{X}_1 \succ \mathcal{X}_2 | s_1, s_2)$ for all observed pairs of actions, while the right panel shows the distribution of the hypothetical probability of winning if the two players are fixed to have equal skill, $\hat{P}(\mathcal{X}_1 \succ \mathcal{X}_2 | 0, 0)$ (the 0 is arbitrary, since only differences in skill enter into the model).

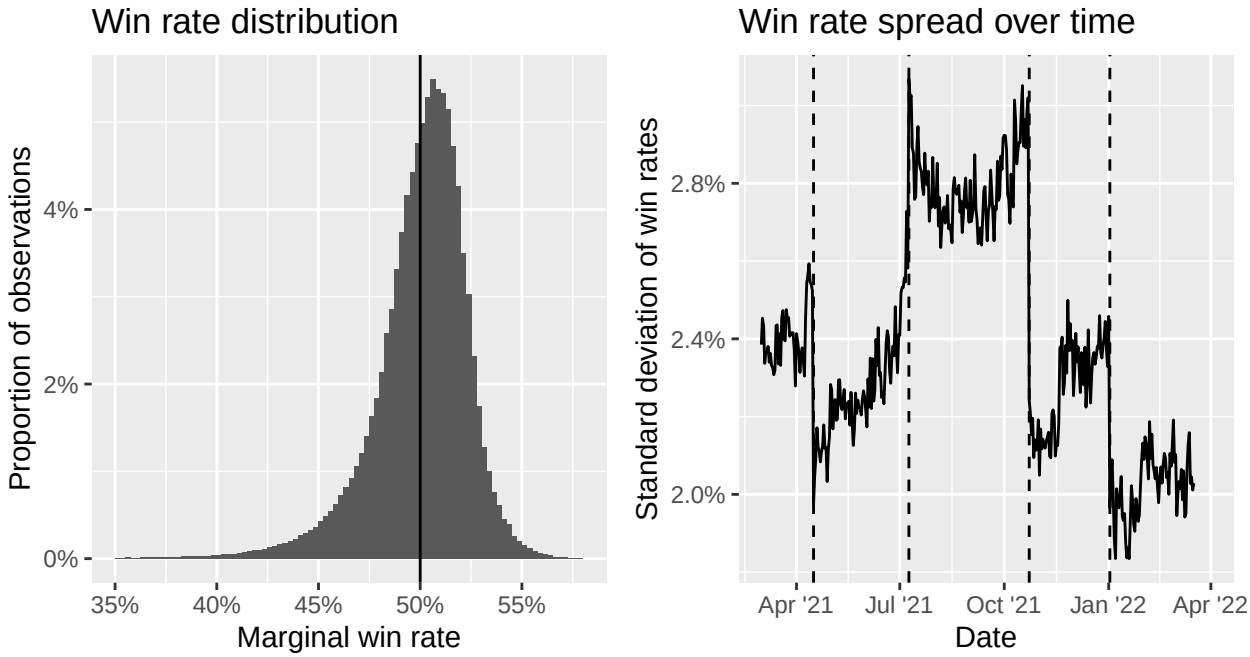
Though the right panel of Figure 11 shows that there is variation in win probabilities between pairs of actions even after controlling for player skill differences, this does not necessarily mean that players are making suboptimal decisions or that the game is not in equilibrium: for instance, in a mixed equilibrium, players may select actions stochastically, with the win probability being above or below 50% for specific pairs of actions, but averaging out to 50% over the distribution of opponent actions. Thus, it is also interesting to consider the overall win rate of an action marginal of the opponent, since this captures the overall normative fitness of an action. If the game is in equilibrium, the marginal win rates of all actions should be 50%.

I calculate the marginal win rate of a given action $\mathcal{X}$ on day t as:

$$\hat{W}_t(\mathcal{X}) = \sum_{\mathcal{X}^{opp}} \hat{P}_t(\mathcal{X} \succ \mathcal{X}^{opp} | 0, 0) \, \hat{P}_t(\mathcal{X}^{opp})$$

i.e., the predicted win probability over an opponent action $\mathcal{X}^{opp}$, assuming equal skill, marginalizing over the empirical distribution of actions on the given day t. This gives a measure of the normative fitness of the action $\mathcal{X}$ under the action distribution on day t. Some summary statistics about the distribution of marginal win rate $\hat{W}_t(\mathcal{X})$ are given in Figure 12. The left panel shows that, in general, the game does not appear to be in equilibrium, and there is substantial arbitrage potential. The distribution is left-skewed, with a long tail of actions with poor win rates (e.g., below 45%). Conversely, very few actions have consistently high win rates (e.g., above 55%). This may suggest

Figure 12: Distribution of model predicted values



Note: in the right panel, vertical dashed lines indicate rule changes.

that, though the game is not in a perfect equilibrium, players are adaptive enough that major arbitrage opportunities (i.e., actions with very high win rates) are not allowed to form. The right panel segments this distribution by day and plots its standard deviation over time. Though there are level shifts at each rule change, the spread does not decrease between rule changes, suggesting that the game does not equilibrate over time.

In Appendix C.3, I show the face validity of the learned embedding space by giving examples of nearest neighbors and a linear interpolation in the embedding space. Though these examples require significant domain knowledge to parse, the nearest neighbor example shows that distances in the embedding space meaningfully capture qualitative strategic similarities between actions, while the linear midpoint between two actions in the space meaningfully interpolates between their strategic properties. It is this smoothness and linearity that allows for the subsequent behavioral analyses in Sections 6.2 and 7.1.

6 Behavioral Analysis of Player Decisions

With the above supervised representation learning model estimated, it is now possible to apply the predictions and embeddings learned from the model towards behavioral analysis of player decision-making. I estimate two reduced form models to understand what weight players put on different information when making two decisions: (1) whether to change their action from one time period to the next and, (2) conditional on making a change, what action to select next. In the following subsections, I discuss the specification, identification, and results of these two models, along with a

discussion and comparison to existing literature.

The goal of this analysis is to understand how players make decisions in a complex competitive environment and how normatively adaptive their decisions are. As discussed in Section 4.3, model-free descriptive analyses can show suggestive patterns, but do not separate out different factors that may be driving player decisions, such as whether players account only for observed outcomes in their decisions or whether they are also able to reason about counterfactual performance of an action against unseen opponents. Incorporating many possible factors jointly into a model allows me to separate out their effects on player decisions. Additionally, by estimating the interaction effects between decision variables and other contextual factors, I assess under what conditions players tend to put more or less weight on different factors..

I note that these models comprise a proof of concept and do not necessarily include an exhaustive set of variables a player may consider when making a decision. Likewise, while I focus on reduced form models, the representations learned from the neural network may also be used towards structural mathematical models of the player decision process to formally test different theories of decision-making, as discussed further in Section 8.

6.1 How Do Players Decide Whether to Change Actions?

6.1.1 Model Specification and Identification

First, I analyze the binary player decision on whether to change their action selection from one time period to the next. I do so via a fixed effects logistic regression of whether player i's action in their t-th hour of play differs from the action used in their $(t-1)$-th hour of play:

$$P\left(\mathcal{X}_{i,t} \neq \mathcal{X}_{i,t-1}\right) = \sigma\left(\alpha_i + \vec{\beta}'\vec{x}_{i,t}\right)$$

where $\vec{x}_{i,t}$ represents a vector of decision variables that the player may take into consideration when making this decision. Subsequently, I refer to this model as the "whether to change" model. I first present a main effects specification including only the decision variables as regressors, while I later include interaction effects with other contextual factors to assess heterogeneity. I consider the following decision variables:

- **Empirical win rate:** the player's empirical proportion of wins, cumulative since their last change, calculated directly from the data with no model. A negative coefficient on this variable means that players reinforce based on directly observed win/loss outcomes, being more likely to change after losing.

- **Skill adjustment:** the player's a priori expected win rate based on skill difference from their opponents, cumulative since their last change, calculated using the skills estimated from the TTT model described in Section 3.3. A positive coefficient means that players are able to adjust for their skill differences from their opponent. For instance, if a player lost to an opponent who is far more skilled than them at playing the in-game match, this loss should

not be attributed to a problem with the player's action selection. Though skill is not directly observable to players, players can observe each other's Elo ratings at the time of the match, providing a signal of the opponent's skill. Players may also be able to infer their opponents' skill based on in-game moves.

- **Stochasticity adjustment:** the player's a priori expected win rate based on their action and opponents' actions, cumulative since their last change, calculated using the fitted values from the proposed neural network after subtracting out skill adjustment. A negative coefficient on this variable means that players reinforce their decisions based on the a priori win probabilities rather than just observed win outcomes. This would mean that players are able to adjust for the stochasticity in the outcome of the game (e.g., if a player has a high a priori chance of winning against a given opponent but loses due to bad luck, this loss should not be attributed to a problem with the player's action selection).

- **Number of games (square root):** the square root of the number of games the player has played, cumulative since their last change, calculated from the data with no model. A positive coefficient indicates that players exhibit variety-seeking behavior as they gain more experience with an action, while a negative coefficient indicates that they exhibit increasing inertia, getting "locked into" the action.

- **Theoretical win rate:** the marginal win probability of the player's current action against a random opponent of the same skill. This is calculated from the proposed neural network by marginalizing the predicted values over the empirical distribution of all actions played on the day containing observation i, t. A negative coefficient on this variable means that players are able to reason about and react to the overall normative fitness of their current action $\mathcal{X}_{i,t-1}$ beyond the specific games played using that action. Though players do not directly observe this quantity, they can obtain further information about the distribution of actions being used by other players (e.g., through past games played with a different action, or by observing other players or consulting outside statistics about the action distribution) and then use model-based reasoning to assess their overall fitness.

Since the first three variables are sample means, their statistical precision scales with the number of games played since the player's last change. Accordingly, I scale these variables (after centering) by the square root of cumulative games played since the player's last change so as to allow the weight placed on these decision variables to scale with their precision. I present results for alternative scalings in Appendix D.1 (linear scaling or no scaling by number of games). Results are largely directionally the same across scalings, but in practice, square root scaling achieves the best model fit. This model makes use of the outputted predictions from my neural network, but does not make use of the learned embeddings.

The use of player fixed effects controls for cross-player heterogeneity, such that the identification of coefficients is not driven by static differences between players that are correlated with their play styles. Rather, identification is driven by within-player variation in actions win/loss outcomes.

Even conditional on player and action, there is still substantial linearly independent variation to identify each coefficient separately. The coefficient on skill adjustment is identified by randomness in matching (sometimes players will match opponents more skilled than them, sometimes less) while the coefficient on number of games played is identified by variation in how long it has been since the player's last change (which is correlated to other decision variables in the model, but there is sufficient residual variation for identification). The coefficient on theoretical win rate is identified by (1) multiple different actions chosen by the same player and (2) shifts in the opponent action distribution over time (e.g., even holding the player action fixed, shifts in the overall action distribution over time will change its normative fitness). The empirical win rate and stochasticity adjustment terms are both essentially noisy signals of the theoretical win rate, but it is this noise which makes their coefficients separately identifiable: the stochasticity adjustment term conditions on the specific opposing actions the player matched against, which will be a random subset of the population action distribution, while the empirical win rate term further conditions on the stochastic game outcome.

6.1.2 Main Effects

In estimating the model, I begin with the subset of data presented in Section 4.2 and further remove players whose fixed effects are not identifiable (i.e., players with no variation in whether they made a change or not). This leaves 5,010,537 observations (player-hour pairs) across 117,409 players.

The results for the main specification are given in Table 3. All coefficients are significant at the 0.1% level with sensible signs. The coefficient on empirical win rate indicates that when a player has a single win (vs. a single loss) they have, on average, $\exp(1.0119) \approx 2.75$ times lower odds of changing their action. Players also react strongly to outside information, with a 1% increase in theoretical win rate corresponding to 4% lower odds of changing. The ratio of these two coefficients indicates that 16 games played with the current action is the approximate equivalence point at which the empirical win rate carries as much weight in the player's decision as the theoretical win rate.

Players are able to adjust to some extent for factors outside their control (their skill difference relative to their opponent, stochasticity in the outcome of the game), but the weights placed on these adjustment factors are more than 90% smaller than the coefficient on the raw win/loss outcomes, indicating that players tend to greatly undercorrect for them. Rather than becoming bored and variety-seeking as they play more games with a given action, players show evidence of inertia wherein they become increasingly locked into their current action as they accumulate experience it. Lastly, the adjusted McFadden pseudo-R^2 of this model is 14.32%, even conditional on fixed effects. This suggests that players' decisions on whether to change actions are highly noisy and/or depend on other contextual variables not accounted for in this model.

All-in-all, these findings show that players are responsive to a variety of performance-related factors in deciding whether to change actions over time. They respond not only to observed outcomes, but also are able to apply outside information about the population action distribution to assess normative fitness. This said, on average they also greatly undercorrect for factors outside their

Table 3: "Whether to change" model estimates (main specification)

Variable	Coefficient (SE)
Empirical win rate	−1.0119 (0.0042)***
Skill adjustment	0.0775 (0.0057)***
Stochasticity adjustment	−0.0764 (0.0159)***
Number of games (square root)	−0.1604 (0.0013)***
Theoretical win rate	−3.9896 (0.0785)***
Observations	5,010,537
Fixed effects	117,409
Overall McFadden R^2	14.32%
Partial McFadden R^2 of regressors	5.92%

Note: ***: $p < 0.001$; **: $p < 0.01$; *: $p < 0.05$; †: $p < 0.1$. Standard errors are clustered at the player level but do not account for first-stage uncertainty from estimation of the neural network and skill models. The first three regressors are scaled by the square root of the number of games played since the previous action change.

control and tend to exhibit inertia wherein they get locked into an action over time.

Note that the normatively optimal decision rule on whether to change actions depends on what decision rule is used to then select an action conditional on a change. If a player knows that they will always be able to perfectly calculate the best response to the population action distribution, then the player should always change actions in anticipation of this (unless their current action is already optimal). More realistically, if a player's action selection is suboptimal and/or noisy, then it will only be optimal to make a change when the expected change in normative fitness is positive. Thus, the fact that players often do not change actions from one hour to the next is not necessarily suboptimal: players may optimally avoid making changes when performing well to avoid regression towards the mean. I explore this possibility in my subsequent normative analysis in Section 7.1.

In Appendix D.1, I show results of alternate specifications such as excluding fixed effects and using a linear probability model instead of logistic regression. I find that, largely, the qualitative results are the same as the specification shown here.

6.1.3 Interaction Effects

In addition to my main effect specification, I also explore heterogeneity in the weights players place on the above decision variables by considering a specification where I include interaction terms with other variables. I interact each decision variable by player skill (calculated from the TTT model), experience (the total number of games the player has played, logged), and time since the last rule change (in hours, logged). This helps reveal insights as to when players place more or less weight on different decision variables.

The results of this specification are given in Table 4. All else equal, skilled players have higher baseline propensity to change and are more attentive to every decision variable, with higher magnitudes on all coefficients. The magnitude of the interaction effects relative to the main effects indicate that skilled players are proportionally more attentive to theoretical win rate and adjustment factors compared to raw empirical win rate. Thus, these players' good normative performance (as captured

Table 4: "Whether to change" model estimates (interacted specification)

Variable	Main effect	Skill interact	Exp. interact	Time interact
Empirical win rate	−1.0060***	−0.1731***	0.0358***	0.0057†
	(0.0039)	(0.0058)	(0.0026)	(0.0029)
Skill adjustment	0.0380***	0.0570***	−0.0345***	−0.0220***
	(0.0055)	(0.0078)	(0.0033)	(0.0042)
Stochasticity adjustment	−0.1653***	−0.0660**	0.0047	0.0448**
	(0.0159)	(0.0247)	(0.0104)	(0.0152)
Number of games (square root)	−0.1360***	0.0758***	−0.0448***	0.0074***
	(0.0012)	(0.0018)	(0.0008)	(0.0009)
Theoretical win rate	−3.9296***	−1.4551***	1.0367***	0.0493
	(0.0848)	(0.1147)	(0.0483)	(0.0680)
Skill	0.2132***			
	(0.0579)			
Experience	−0.3744***			
	(0.0240)			
Time since rule change	−0.0986**			
	(0.0339)			
Observations	5,010,537			
Fixed effects	117,409			
Overall McFadden R^2	14.59%			
Partial McFadden R^2 of all regressors	6.24%			
Partial McFadden R^2 of interactions	0.33%			

Note: ***: $p < 0.001$; **: $p < 0.01$; *: $p < 0.05$; †: $p < 0.1$. Standard errors are clustered at the player level but do not account for first-stage uncertainty from estimation of the neural network and skill models. The first three regressors are scaled by the square root of the number of games played since the previous action change. All variables are centered before interacting, such that the main effects correspond to the coefficient for an observation with average skill, experience, and time since rule change. Experience indicates the total cumulative number of games a player has ever played (logged) while time since rule change indicates the number of hours since a rule change took place (logged).

by skill) may in part be due to them being better at reasoning about variables not directly observed.

Conversely, holding skill constant, more experienced players are proportionally *less* attentive to factors like skill adjustment and theoretical win rate. Players also correct for stochasticity and skill differences less as time passes after a rule change. The results on number of games played indicates that, while players on average get "locked into" actions over time, higher skill and longer time since a rule change correspond to more variety-seeking, while more experience corresponds to more lock-in.

Players also appear to become less attentive to decision variables as they accumulate experience (holding skill constant) and as time passes since a rule change, resulting in noisier decisions. Players who perform normatively better (as captured by skill) tend to be more attentive overall, and in particular place higher weight on outside information and adjustment factors while also being less prone to lock-in; thus, there is substantial heterogeneity in how well players are able to account for information not directly observed, with some players being highly responsive to these factors, while others appear to make decisions largely based on raw observed outcomes.

6.1.4 Comparison to Previous Findings

Though the complexity of the task and numerous variables involved in the decision makes these findings difficult to compare directly to existing literature, the findings are qualitatively consistent with those of Khaw et al. (2017), who have participants engage in a repeated probability estimation task. In this task, participants repeatedly observe draws (with replacement) of red or green rings from a box containing an unknown proportion of red vs. green rings and attempt to estimate the true proportion, with an opportunity to update their estimate after each observation. The experiment is incentive compatible, with the expected payout to the participant decreasing in the squared difference between the estimated proportion and true proportion. At each trial, there is a small probability that the box is replaced by a new box with a different true proportion. The participant knows the probability of a change occurring, but does not directly observe whether a change has occurred at any given time.

This task mirrors the structure of the decision problem in my empirical context, wherein players' win outcomes depend on an imperfectly observed state variable (the action distribution of other players) which is evolving over time. But, their task is much simpler since the decision is scalar and depends only on a single scalar unobserved state variable. Thus, comparing results can provide suggestive evidence as to how their results generalize to more complex tasks where there is more information for players to process.

Khaw et al. (2017) find that participants are "sticky," only adjusting their probability estimates in 8.9% of observations. I find that players change actions in 34.0% of hourly observations, and each hour contains an average of 3.96 matches, suggesting an overall change probability of approximately 8.6%.[15] Thus, the overall frequency of change is comparable. They further find that the logit probability of adjustment is linearly increasing in the squared difference between the current estimate and optimal Bayesian estimate, with a McFadden R^2 of 2.2%. In my setting, there are multiple sources of information (directly observed performance, contextual factors such as skill and stochasticity, overall theoretical performance) which all play a role in player decisions and in total explain a larger proportion of variation in decisions (partial McFadden R^2 of 5.92% for all regressors; 3.07% for empirical win rate).

Thus, the decision of whether to change actions does not appear to be any noisier in this complex environment than in the simpler setting of Khaw et al. (2017). This may be due to a couple of factors: first, though the action space is much more complex, the decision of whether to change is still binary, and the decision variables are scalar, so the cognitive effort required to make the decision is likely dramatically different; second, players in my context tend to be highly motivated and have substantial experience playing the game, and so are likely to be quite engaged and attentive, whereas the participants in Khaw et al. (2017) engage in a relatively monotonous standalone task where the potential gain from playing optimally (relative to naively guessing 0.5 probability at every trial) is \$1.60 per 1,000-trial session.

[15] Likely a slight underestimate, since the hourly aggregation smooths over transitory changes that happen within an hour.

Thus, though players appear prone to certain biases (e.g., undercorrecting for contextual factors and getting locked into actions), their decisions of whether to adjust actions from one period to the next in this complex game are just as attentive as (if not more than) the decisions of participants in a simpler but low-stakes stylized lab task.

6.2 How Do Players Select New Actions?

6.2.1 Model Specification and Identification

Having analyzed the variables driving players' decisions on whether to change actions over time, I next analyze what factors drive players' action selection conditional on making a change. This is where the embeddings learned by my supervised representation learning model come into play. I model the embedding of the new selected action $\mathbf{z}_{i,t} = \hat{f}(\mathcal{X}_{i,t})$ as a fixed effects vector linear model:

$$\mathbf{z}_{i,t} = \boldsymbol{\alpha}_{i,r(t)} + \sum_p \beta_p \mathbf{z}_{i,t}^{(p)} + \boldsymbol{\varepsilon}_{i,t}$$

where $\mathbf{z}_{i,t}^{(p)}$ is an embedding vector representing a decision rule that a player may consider in their decision and $r(t)$ is the ruleset in effect at time t, such that $\boldsymbol{\alpha}_{i,r(t)}$ is a (vector-valued) player-ruleset fixed effect. Subsequently, I refer to this model as the "how to change" model. I present robustness checks for alternative parameterizations in Appendix D.2. I consider the following variables, which correspond to different decision rules that the player may employ:

- **Own previous action:** the embedding of the action $\mathbf{z}_{i,t-1}$ that player i used previously. A positive coefficient on this variable means that players exhibit inertia, tending to choose actions that are strategically similar to their current action.
- **Action of opponents won (lost) against:** the average of the embedding vectors of actions used by opponents the player won (lost) against, cumulative since the previous change. Positive coefficients on these variables mean that players tend to imitate their recent opponents by choosing actions that share strategic properties with the actions used by opponents. Negative coefficients mean that players tend to avoid actions used by opponents.
- **Population action distribution:** the average of the embedding vectors of actions used by all players on the day before observation i, t. A positive coefficient on this variable means that players use outside information to monitor which actions are popular overall and tend to imitate these actions. The distribution for the preceding date is used to avoid tautological results from the selected action being contained in the action distribution. Though a single observation of $\mathbf{z}_{i,t}$ only contributes a small amount to the overall distribution, it could have outsized influence on the action distribution residual of the fixed effects, especially if other players imitate this action later in the day. Using the lagged action distribution avoids circular results, and may also better capture the information available to the player, since players likely will not have access to real-time updates on the population action distribution.

- **Best response to opponents won (lost) against:** the embedding vector of the best response action that maximizes the average probability of winning against opponents the player won (lost) against, cumulative since the last change. This captures the extent to which players reason counterfactually about how they could improve their performance by choosing actions that would have performed better in recent games (especially for losses; for opponents won against, players may seek to improve the margin of how "safe" the win was).

- **Best response to population action distribution:** the embedding vector of the best response action that maximizes the marginal probability of winning against a random opponent based on the population action distribution on the date preceding time t. This captures the extent to which players use outside information to optimize their actions based on its strategic matchup to the population action distribution as a whole. As with the population action distribution, I lag the date to avoid circular results.

To calculate best responses in practice, I rank order all unique actions observed under the current ruleset and average the embeddings across the 10 actions with the highest win probabilities. This is for two practical reasons. First, though best responses can in principle be solved (approximately)[16] by a linear program, the learned distribution of embeddings may not have support over the entire space of embeddings possible under the model specification (here, a unit hypercube), such that the estimated optimum may not correspond to a legal action. Second, there may be multiple actions that obtain near-optimal performance, and which action happens to be the single highest rank may be an artifact of estimation error. A player who selects an action that is close to any of these near-optimal actions still shows evidence of reasoning about best responses. Selecting the best responses from the actions observed in the data addresses the first issue by ensuring the best response stays within the support of the data, while averaging over the top 10 best responses addresses the second issue by smoothing over outliers and capturing multiple near-optimal solutions. In Appendix D.2, I present robustness checks averaging over alternate numbers of best responses besides 10 and find that results are qualitatively consistent.

The variables split up by whether the player won or lost against the opponent are not always well-defined, since players sometimes have only wins or only losses since their previous action change. I mean impute these variables within each fixed effect level (i.e., within each player-ruleset pair), which is equivalent to omitting these terms for missing observations, so that they do not influence the model output when missing.

The use of player-ruleset fixed effects controls for player heterogeneity in baseline preferences for different types of actions (e.g., due to a player's specific play style), and that those baseline preferences may differ by ruleset, such that decision rules are identified using within-player, within-ruleset variation. The autoregression/inertia term (own previous action) is identified by player-ruleset pairs with at least 3 observations, for whom there is still residual variation in the first-

[16] When calculating the marginal win probability over multiple opponents, the inverse logit introduces a nonlinearity. However, given that most pairwise win probabilities are near 50%, where the inverse logit function is approximately linear, in practice I find that a linear approximation of the objective function results in an almost exact solution.

order autoregressive term after adjusting out fixed effects, such that I can estimate the extent of state-dependence in excess of baseline preferences. The coefficients on opponent actions and their corresponding best responses are identified by randomness in the opponents with which a player matched in their recent play history. The population action distribution and corresponding best response are shared across all players on a given day, so these are identified by aggregate time series variation in the action distribution over time within a ruleset.

A normatively optimal player with full information would place 100% weight on the best response to the population action distribution. The population action distribution may not be perfectly observed to the player, but even so, a normatively optimal player with limited information could still come close to the full information optimum by best responding to their observed opponents (who are random samples from the overall action distribution). This said, calculating best responses requires model-based counterfactual reasoning and is likely to be cognitively difficult for players. Thus, players may also rely on simple model-free heuristics such as imitating their opponents. Additionally, players may follow other idiosyncratic decision rules or be imprecise in their calculations, resulting in deviation from the subspace of embeddings spanned by the decision rules in my model. My model disentangles what decision rules players may follow when selecting actions, including the extent to which players deviate from the decision rules considered (via the error term).

I first consider the main effects of avegage weights placed on each decision rule, followed by a specification including interaction effects to understand when players place more/less consideration on different factors. My subsequent normative analysis in Section 7.1 further considers how normatively well-adapted player decision rules are.

6.2.2 Main Effects

In estimating the model, I begin with the subset of data presented in Section 4.2, subset to player-hour pairs where an action change occurred, remove the first day of each ruleset (due to the aforementioned lagging of the action distribution variables), and remove the first action change of each player-ruleset pair (where the variables in my model are from a different ruleset, and thus not applicable to the decision). I then remove players whose fixed effects are not identifiable (i.e., players with only a single observed change). This leaves 1,699,521 observations (player-hour pairs) across 99,444 players (144,850 player-ruleset pairs). The dependent variable, regressors, and fixed effects are all 100-dimensional variables in the embedding space. I estimate the model via ordinary least squares, equally weighting the 100 dimensions of the dependent variable.[17]

The results for the main specification are given in Table 5. All coefficients are significantly positive at the 0.1% level, except for the best response to the population action distribution, which has a negative coefficient. On average, players place 7.82% weight on their own previous action, demonstrating inertia/preference to stay close to their current action even beyond baseline preferences. Players also tend to imitate the opponents, and as is to be expected, tend to imitate the opponents

[17]This does not account for unequal variances or correlations between dimensions of $\boldsymbol{\varepsilon}_{i,t}$. Better statistical efficiency may be obtained by using generalized least squares, allowing for $\boldsymbol{\varepsilon}_{i,t}$ to have an arbitrary covariance matrix, as in seemingly unrelated regressions.

Table 5: "How to change" model estimates (main specification)

Variable	Coefficient (SE)
Own previous action	0.0782 (0.0015)***
Action of opponents won against	0.0017 (0.0003)***
Action of opponents lost to	0.0084 (0.0003)***
Population action distribution	0.5579 (0.0090)***
Best response to opponents won against	0.0033 (0.0002)***
Best response to opponents lost to	0.0008 (0.0002)***
Best response to population action distribution	−0.0068 (0.0017)***
Observations	1,699,521
Fixed effects	144,850
Overall R^2	19.95%
Partial R^2 of regressors	0.66%

Note: ***: $p < 0.001$; **: $p < 0.01$; *: $p < 0.05$; †: $p < 0.1$. Standard errors are clustered at the player level but do not account for first-stage uncertainty from estimation of the neural network and skill models. R^2 is calculated by pooling residual variance across the 100 dimensions of the dependent variable relative to a null model with ruleset fixed effects.

they lost to (0.84% weight) more so than the ones they won against (0.17% weight). However, by far the largest weight is on the population action distribution (55.8%), indicating that players have a strong tendency to follow aggregate trends in the action distribution, imitating what is popular with other players. Thus, rather than the specific opponents they recently won or lost against, it appears that players are most prone to following overall popular trends.

The best response coefficients further indicate that players are able to reason counterfactually about what actions would have performed better against their opponents, but the weight placed on these decision variables is small; curiously, the weight placed on best responding to the opponents the player already won against (0.33%) is higher than that placed on responding to opponents lost against (0.08%). This, paired with the higher weight placed on imitating opponents lost to, may indicate that players follow different heuristics depending on match outcome. After losing to an opponent, a player may choose to simply imitate them, assuming that the opponent's action is normatively better than their own current action, without needing to do the relatively complex calculation of how to best respond. Conversely, after winning against an opponent, imitation is unlikely to further help, so a player may wish to iterate on their action to make their win even "safer" (i.e., increase the a priori probability of winning).

Most surprisingly, the coefficient on the best response to the population action distribution is negative (−0.68%), indicating that after controlling for other decision variables, players actually tend to move *away* from the overall best response. For comparison, an oracle decision-maker seeking to maximize overall win rate would put 100% weight on this variable. However, this decision rule is also the most difficult for players to calculate since it is two layers separated from what is directly observable to the player: the population action distribution is not directly observed by the player, and best responding to it requires reasoning counterfactually about the performance of an action marginalized over the entire population distribution of actions. Averaging win probabilities over

all possible opponents for each possible action and identifying the best possible action is a very difficult problem for a player to solve, even approximately, especially since Figure 12 shows that the differences in marginal win rates between actions tend to be fairly small. Thus, it may be that reasoning about the population best response is simply too difficult for most players. The negative coefficient may be an artifact of players "overfitting" on other decision variables (e.g., imitating the overall action distribution or best responding to their opponents) such that, after controlling for all other variables, they tend to move away from the overall best response. Note that this does not mean that players necessarily move towards normatively worse actions on average: the other decision variables may also move them towards normatively better actions. The negative coefficient instead reflects that, *in excess of* all other factors, players tend to place slight negative weight on the best response.

Additionally, it is clear that player decisions are driven to a large extent by noise or other idiosyncratic factors not included in the model. The overall R^2 of the model, inclusive of fixed effects, is just below 20%, indicating that the vast majority of variation in action selection is driven by factors other than the main effects of the considered decision rules and baseline player preferences.

6.2.3 Interaction Effects

In addition to my main effects specification, I also explore heterogeneity in the weights players place on the above decision rules by considering a specification where I include interaction terms with other variables. Specifically, I include interactions with three variables from the "whether to change" model (number of games played, empirical win rate, and theoretical win rate) along with the same three interactions included in that model (player skill, experience, and time since the last rule change). This shows the extent to which players employ different decision rules depending on the circumstances.

The results of this specification are given in Table 4. The interaction effects capture substantial additional variation in player action selection. As a player has been using their current action for longer, they place more weight on imitating opponents lost to and best-responding to opponents won against, while placing less weight on the population action distribution. Thus, when players have substantial experience with their current action, they make decisions based more on observed opponents as opposed to the action distribution as a whole. As a player accumulates more overall experience (beyond the current action), they demonstrate more inertia and are less responsive to the population action distribution (and its best response). Additionally, players with more experience are less likely to imitate opponents. As time passes since a rule change, players tend to follow popular actions more, placing less weight on inertia and imitating opponents.

When a player's empirical win rate has been high with their current action (after controlling for the action's theoretical win rate), they exhibit more inertia, are less prone to imitation, and place more weight on best responding to the opponents they won against. Conversely, when a player's theoretical win rate is high, they exhibit less inertia, place more weight on the action distribution and its best response, and tend to imitate opponents lost to (rather than best responding to them).

Table 6: "How to change" model estimates (interacted specification)

Variable	Main effect	# of games (sq. root)	Empirical win rate	Theoretical win rate	Skill	Experience	Time since rule change
Own previous action	0.0721***	0.0002	0.0407***	−0.7275***	−0.0165***	0.0402***	−0.0254***
	(0.0013)	(0.0004)	(0.0016)	(0.0229)	(0.0018)	(0.0009)	(0.0007)
Action of opponents won against	0.0008*	−0.0005†	−0.0008	−0.0164	0.0009†	−0.0006**	−0.0010***
	(0.0003)	(0.0003)	(0.0011)	(0.0104)	(0.0005)	(0.0002)	(0.0003)
Action of opponents lost to	0.0102***	0.0042***	−0.0116***	0.0296**	0.0012*	−0.0007***	−0.0016***
	(0.0004)	(0.0003)	(0.0011)	(0.0105)	(0.0005)	(0.0002)	(0.0003)
Population action distribution	0.5929***	−0.0036***	−0.0260***	0.6807***	−0.0460***	−0.0391***	0.0293***
	(0.0085)	(0.0006)	(0.0025)	(0.0327)	(0.0029)	(0.0010)	(0.0010)
Best response to opponents won against	0.0030***	0.0005***	0.0036***	0.0046	−0.0011**	0.0002	−0.0004
	(0.0002)	(0.0002)	(0.0009)	(0.0088)	(0.0004)	(0.0001)	(0.0002)
Best response to opponents lost to	0.0007***	0.0001	−0.0003	−0.0292***	0.0008**	0.0000	0.0003
	(0.0002)	(0.0001)	(0.0008)	(0.0077)	(0.0003)	(0.0001)	(0.0002)
Best response to population action distribution	−0.0073***	−0.0003	0.0006	0.0975***	0.0108***	−0.0014**	−0.0008†
	(0.0017)	(0.0002)	(0.0008)	(0.0125)	(0.0020)	(0.0004)	(0.0004)
Observations	1,699,521						
Fixed effects	144,850						
Overall R^2	20.31%						
Partial R^2 of all regressors	1.11%						
Partial R^2 of interactions	0.45%						

Note: ***: $p < 0.001$; **: $p < 0.01$; *: $p < 0.05$; †: $p < 0.1$. Standard errors are clustered at the player level but do not account for first-stage uncertainty from estimation of the neural network and skill models. R^2 is calculated by pooling residual variance across the 100 dimensions of the dependent variable relative to a null model with ruleset fixed effects. All variables are centered before interacting, such that the main effects correspond to the coefficient for an observation with all other variables fixed at their averages.

Thus, when players already have a normatively good action, they focus on iterating to make it better by incorporating elements of the actions of opponents they lost to, popular actions overall, and best responses to the overall action distribution. However, if a player had a high empirical win rate simply by getting lucky (rather than having a normatively good action), the patterns are largely the opposite. The fact that players place less weight on imitation after having a high empirical win rate is consistent with the findings shown in Figure 7.

All else equal, players with higher skill are less prone to inertia and place more weight on imitating and best responding to opponents lost to. They also tend to focus on best responding to the population action distribution rather than imitating it.

Overall, these results show that the weights players place on different decision rules can vary greatly by context. Most notably, the interaction effects for best response to the population action distribution indicate that even if players on average place slightly negative weight on this factor, players do sometimes place positive weight on the population best response, for instance when they are highly skilled and/or their current action is already normatively good. I consider in Section 7.1 whether these patterns are adaptive, i.e., whether players place more weight on decision rules when they are likely to perform better.

6.2.4 Comparison to Previous Findings

As with the "whether to change" model, the complexity of my empirical setting makes parameter estimates difficult to directly compare to existing literature. However, interesting comparisons can be drawn to findings based on the experience weighted attraction (EWA) model (Camerer and Ho, 1999). In the EWA model, players select actions in proportion to learned payoffs, observed or counterfactual. In particular, a player maintains "attractions" towards each action and updates their attractions after each game in proportion to the payoff the action would have had against the observed opponent action, but possibly places more weight on the observed payoff (i.e., the outcome for the selected action) than the counterfactual payoffs (i.e., the foregone outcomes of unselected actions). At the next game, the player selects their action based on a stochastic decision rule (usually multinomial logit for a discrete action space) as a function of the attractions. The EWA model includes a weight parameter δ which captures how much weight players place on counterfactual outcomes compared to the observed outcome when updating attractions. When $\delta = 1$, players fully internalize counterfactual outcomes, and so will tend to update their actions towards best responses. Conversely, when $\delta = 0$, players only account for directly observed outcomes, corresponding to pure reinforcement learning. In a review of EWA model estimates from 31 datasets covering 20 unique games, Camerer et al. (2002) find that δ is generally between 0.5 and 1 for games that admit pure strategy equilibria, while δ tends to be close to zero for games that admit only mixed strategy equilibria.

In my context, due to the complex action space, instead of modeling the probabilities of selecting each individual action, I model the selected action as a linear combination of candidate actions in the embedding space, and as such, the coefficients from my model are not directly comparable

to estimates from an EWA model. Nonetheless, I can compare the relative weight players place on actions that require counterfactual reasoning and those that can be based directly on observed outcomes and contrast to the δ parameter of the EWA model. In particular, comparing the weights placed on imitating vs. best responding to opponents, players on average place a combined weight of 0.0101 on imitation as compared to 0.0041 on best response; thus, players on average place 2.5 times as much weight on decision rules that can be obtained by model-free imitation as on decision rules that require model-based counterfactual reasoning. Though not directly comparable to EWA parameters, this finding is similar to the finding in the EWA literature that players place 1-2 times as much weight on observed outcomes as foregone counterfactual outcomes when updating attractions in games that admit pure strategy equilibria. Conversely, when comparing the weight placed on the population action distribution as opposed to its best response, I find that players put orders of magnitude more weight on imitation than best response, similar to EWA findings that players place little to no weight on counterfactual outcomes in games that only admit mixed strategy equilibria.

As discussed above, calculating the best response (or generally finding actions that are normatively well-adapted) to the population action distribution is extremely difficult for the player since it involves taking an average over a large number of possible opponents (whose distribution is not directly observed) and distinguishing between actions that have very similar win rates; conversely, calculating best responses against specific opponents will be more feasible since this does not require averaging over many opponents and there will generally be more variability in win probabilities; thus, the fact that players do put positive weight on best responses to the specific opponents they matched against—but put little (or even negative) weight on the best response to the population action distribution—may be explained by the relative difficulty of identifying the best response in one case versus the other. This mirrors a key feature of games that only admit mixed strategy equilibria, which is that when the population action distribution is near equilibrium, all outcomes will have almost identical expected payoffs, making it difficult for players to differentiate between actions and converge on a best response (as discussed by Camerer et al., 2002).

Thus, a common finding between this literature and my empirical context is that when it is relatively easy to differentiate between the expected payoffs of actions (e.g., because the payoffs differ significantly and/or the expectation is easier to calculate due to be taken over relatively few opposing actions), players can successfully reason about counterfactuals and best responses to an extent, but when this calculation is too difficult, players fall back on model-free learning tactics. All-in-all, even with a much larger action space and complex payoff structure, players still show evidence of counterfactual reasoning to a comparable magnitude to that found in stylized lab studies, with analogous boundary conditions. This may in part be reflective of players being highly motivated and learning over a longer timescale, such that they have more opportunity and incentive to perform sophisticated reasoning, offsetting the added complexity of the decision problem relative to the simpler games studied in prior literature.

7 Normative and Managerial Implications

7.1 Normative Properties of Player Behavior

As seen earlier in Figure 12, there is nontrivial dispersion in the normative fitness (theoretical win rates) of actions, and this dispersion does not appear to dissipate over time. Section 6.2 further showed that players tend to rely heavily on heuristics such as imitation rather than calculating counterfactual best responses. These findings raise the question of whether player decisions tend to lead to normatively better actions (i.e., improvement in theoretical win rate compared to their previous action) and what implications player decision rules have for the equilibration (or lack thereof) of the game.

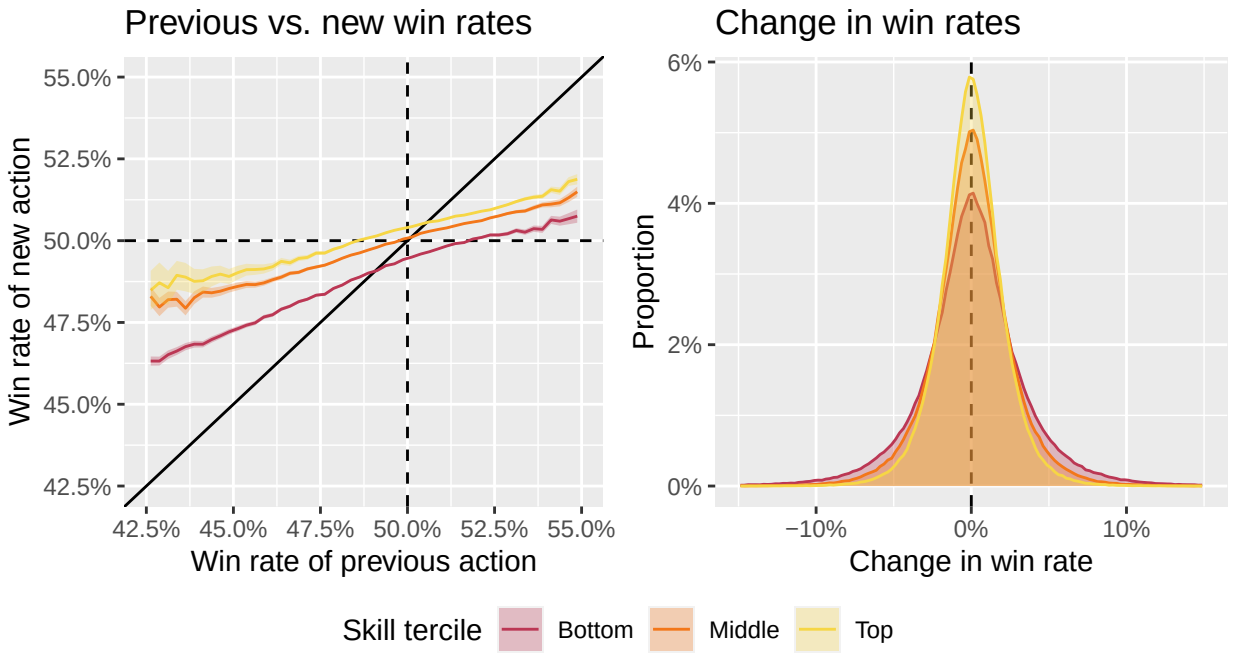
Figure 13 shows how normatively adaptive player action selection tends to be. The left panel shows the average relationship between theoretical win rate before and after an action change, segmented by skill, while the right panel shows the distribution of change in win rate (before vs. after change). Substantial regression towards the mean is evident: when players' previous action was bad, they tend to choose better actions, and vice-versa. The average change in win rates is only slightly positive, with large noise around that average: the average improvement in win rate is 0.04% compared to standard deviation of 2.89%, with only a 50.75% probability of improvement. This same qualitative pattern holds across skill levels, though higher skilled players do, on average, select better actions with lower variance. Thus, player action selection appears to be highly noisy; this may explain why the game does not appear to equilibrate over time, since the regression to the mean effect means that changing actions repeatedly does not necessarily bring a player close to the population best response.

This finding raises the question: what aspects of players' decision rules make them suboptimal, and what decision rules lead to better/worse normative performance? Do players anticipate their own suboptimality in action selection when deciding whether to change actions? For instance, if players are aware of when they are more or less likely to make a good decision, they may adaptively choose to only make changes when they are relatively likely to make a good decision. To address these questions, I run supplementary analyses that consider under what conditions players tend to make normatively better/worse decisions. I have two such analyses: one assessing normative fitness of player decisions as a function of contextual factors, then an extension modeling fitness as a function of decision weights.

7.1.1 Normative Fitness as a Function of Contextual Factors

First, Table 7 shows estimated coefficients for a fixed effects linear regression regressing the theoretical win rate of a player's selected action on the six interaction variables considered in Table 6 along with player fixed effects. For readability of the table, theoretical win rates (both the dependent variable of the new action and regressor of the previous action) are multiplied by 100 to be on a percentage point scale. Consistent with Figure 13, the win rate of the previous action is positively correlated with the win rate of the newly selected action, but the slope is substantially less than 1,

Figure 13: Normative fitness of action selection



Note: the bands in the left panel represent pointwise 95% confidence intervals while the black line represents a unit slope relationship with no regression towards the mean. Both plots are calculated by discretizing their respective x-axes to bins of 0.25% width.

indicating a regression to the mean effect. Additionally, players' new action selection tends to be worse when they won more matches with their previous action and when they had more experience playing with that action (and more experience overall). Conversely, player action selection tends to be better when the player is more skilled and when more time has passed since a rule change.

The finding that skilled players select normatively better actions is unsurprising, since skilled is defined in terms of player performance. The regression to the mean effect can arise from player inertia and noise in the decision-making process. Since players exhibit inertia, they will tend to choose better actions when their starting place is already good. But, as seen earlier, a large proportion of variance in player decisions is due to noise or idiosyncratic factors, which can lead to regression to the mean. Conversely, after good empirical performance (holding the theoretical win rate constant), players tend to make worse decisions, perhaps reflecting players becoming overconfident after a streak of good luck.

As time passes since a rule change, players also likely become more familiarized with the current ruleset, allowing them to make better decisions. As players have used their current action for longer, they tend to make worse decisions, which could mean that players become "rusty" at selecting actions when it has been a long time since they last made a selection. Players also tend to make worse decisions as they accumulate more experience overall (holding skill constant), which could reflect players becoming bored or inattentive when they have played the game for a long time.

Comparing the results of this model to the main effects of the "whether to change" model from Table 4, the coefficient signs largely match: under conditions where players tend to make normatively

Table 7: Model of win rate of selected action (contextual factors only)

Variable	Coefficient (SE)
Number of games played with previous action (square root)	−0.0129 (0.0012)***
Empirical win rate of previous action	−0.2242 (0.0072)***
Theoretical win rate of previous action	0.1362 (0.0019)***
Skill	1.5078 (0.0169)***
Experience	−0.0446 (0.0031)***
Time since rule change	0.0471 (0.0034)***
Observations	1,699,521
Fixed effects	99,444
Overall R^2	28.44%
Partial R^2 of regressors	3.99%

Note: ***: $p < 0.001$; **: $p < 0.01$; *: $p < 0.05$; †: $p < 0.1$. Standard errors are clustered at the player level but do not account for first-stage uncertainty from estimation of the neural network and skill models. For readability, the dependent variable and "theoretical win rate of previous action" are both multiplied by 100 (percentage point scale).

worse decisions, they are also less likely to decide to make a change at all.[18] Thus, these findings are consistent with players being adaptive in their choice of whether to change actions, anticipating the conditions under which they are more/less likely to make a good normative change and selectively only changing actions when they are relatively likely to make a positive change.

7.1.2 Normative Fitness as a Function of Decision Rules

Next, I consider not only how contextual factors correlate to the normative fitness of the selected action, but also the normative fitness of different decision rules. Since each action selection is a point in a 100-dimensional space, it is possible to solve for a vector of decision weights $\boldsymbol{\beta}_{i,t}$ for each individual observation and utilize variation in the decision weights to understand when different heuristics leads to normatively better/worse decisions. Hypothetically, if players always made decisions optimally, then these decision weights would all be uncorrelated with normative fitness, since the decision weights would always satisfy the first order condition with respect to win rates. Conversely, a negative (positive) correlation between a decision weight and normative fitness would indicate that players over- (under-)weight a decision rule and could improve performance by weighing that factor less (more) heavily in their decisions.

Of course, observation-level estimates of $\boldsymbol{\beta}_{i,t}$ are likely to be highly noisy due to collinearity between the decision rules and due to the added imprecision from centering out the fixed effects. As such, for this analysis, I subset to player-ruleset pairs with at least 10 observations so that the fixed effects can be estimated relatively precisely, and further add a small ridge penalty when solving for the decision weights so as to ensure numerical stability.

Specifically, denoting the embedding of the action selected by player i at time t after subtracting out fixed effects as $\tilde{\mathbf{z}}_{i,t}$ (a 100-dimensional vector) and the matrix of decision rule action embeddings

[18] The theoretical win rate of the new action is positively correlated with that of the previous action, but since the slope of the relationship is less than one, higher win rate corresponds to lower *change* in win rate, so the negative coefficient on current theoretical win rate in Table 4 is consistent with this pattern.

(after centering by fixed effects) as $\tilde{\mathbf{X}}_{i,t}$ (a 100-by-7 matrix), I estimate the observation-specific vector of decision weights as:

$$\hat{\boldsymbol{\beta}}_{i,t} = \left(\tilde{\mathbf{X}}'_{i,t}\tilde{\mathbf{X}}_{i,t} + \lambda\mathbf{I}\right)^{-1}\tilde{\mathbf{X}}'_{i,t}\tilde{\mathbf{z}}_{i,t}$$

where $\mathbf{I}$ is the 7-by-7 identity matrix and λ is the ridge penalty factor (for the results below, I use $\lambda = 0.01$). The vector $\hat{\boldsymbol{\beta}}_{i,t}$ gives an estimate of how much weight the player put on each decision rule (e.g., imitating opponents lost to, best responding to opponents won against) when selecting their action. These estimates are likely to be fairly noisy, but variation in the weights across observations still allows for directional insights into when each decision rule performs normatively well/poorly.

Additionally, these decision rules will vary across observations in how well they explain the chosen action: sometimes player actions may be almost perfectly explained as a linear combination of the decision rules considered in my model, while at other times the player action selection may not correspond closely to any linear combination of decision rules considered. To measure the extent to which player actions deviate from the space spanned by the decision rules considered, I calculate the *action embedding deviation* as the Euclidean distance between the selected action and its predicted value, i.e., the L_2 norm of the residual:

$$\hat{e}_{i,t} = \sqrt{\left(\tilde{\mathbf{z}}_{i,t} - \tilde{\mathbf{X}}_{i,t}\hat{\boldsymbol{\beta}}_{i,t}\right)'\left(\tilde{\mathbf{z}}_{i,t} - \tilde{\mathbf{X}}_{i,t}\hat{\boldsymbol{\beta}}_{i,t}\right)}$$

This captures the extent to which player decisions deviate from what can be explained by the considered decision rules, either due to noise in the decision-making process or other idiosyncratic considerations the player takes into account when selecting an action. Note that a large deviation is *not necessarily* normatively bad: while pure decision noise will tend to make a player worse off in expectation, if the deviation is due to a player identifying an action outside the space of considered decision rules that yields normative improvement, the deviation may not be bad.

Table 8 shows the estimates of a fixed effects linear model of the theoretical win rate of the selected action as a function of the decision weights and error, interacted with the six contextual factors considered above, with player fixed effects. The main effects show that, on average, inertia and imitation of opponents (especially opponents won against) leads to worse normative decisions, while placing more weight on best responding to opponents lost to and the population action distribution as a whole lead to better decisions. There is a strong negative effect of action error, indicating that noise and/or other idiosyncratic factors in player decisions besides the considered decision rules tend to lead to worse normative performance.

Thus, though relying on model-free heuristics may not always be bad, these results suggest that on average, players exhibit too much inertia and overrely on simple imitation of opponents. Likewise, players could do better by putting more weight into best responding to the opponents they lost to and best responding to the action distribution as a whole. Additionally, staying closer to the space of linear combinations of the considered decision rules, e.g., by reducing noise in the decision calculation or relying less on idiosyncratic factors, can greatly improve player performance. The main effects of imitating the population action distribution and best responding to opponents

won against are null. This may indicate that, on average, players allocate a near-optimal degree of weight to these factors.

Beyond the main effects, it is interesting to consider how the normative properties of different decision rules vary as a function of contextual variables. For instance, the main effects show that, on average, deviating from a linear combination of the considered decision rules leads to worse normative performance, but it may be the case that players are sometimes able to identify favorable deviations. The interaction effects show under what conditions different decision rules can lead to better or worse performance.

When a player has more experience with their current action, the negative effects of inertia and imitation are mitigated, while the effects of best responding to opponents and effects of error are amplified. This is intuitive, since when a player has more experience with an action, the opponents played against will be a more reliable signal of the action distribution as a whole, giving players more useful information off of which to make decisions. Conversely, as a player accumulates more experience overall, this yields a positive effect of imitating the action distribution at large and mitigates the negative effects of imitating and responding to opponents won against, perhaps indicating that experienced players are better able to discriminate which aspects of other players' actions (especially those won against and the population distribution at large) are normatively useful to incorporate and which should be ignored. As time passes after a rule change, it appears that paying attention to the population action distribution becomes less important, and action errors are less detrimental.

When a player's previous action was already normatively good (high theoretical win rate), inertia is beneficial and the negative effects of error and population imitation are amplified, while the negative effects of imitating opponents are mitigated. When the empirical win rate is high, similar patterns hold, but imitating and best responding to opponents lost to are less favorable, likely reflecting that the sample size of opponents lost to will be smaller and less reliable when the empirical win rate is high. Lastly, skilled players have an amplified negative effect of inertia and mitigated effects of imitating and best responding to opponents. Most notably, there is a strong positive interaction between action deviation and skill. Taken together, these results indicate that the normative performance of skilled players' decisions are less sensitive to which decision rules they follow, likely indicating that their decision weights are closer to optimal, except for an over-reliance on inertia. Additionally, skilled players are better able to identify favorable deviations from the considered decision rules.

In sum, there are a number of factors that lead to player action selection being normatively suboptimal. Players are over-reliant on heuristics based on model-free, directly observed information including inertia and imitation of opponents, while failing to consider counterfactual best responses that are more difficult to compute but could lead to better decisions. Noise and/or idiosyncratic factors driving player decisions also result in normatively worse decisions. Skilled players, who on average make normatively better decisions, achieve better performance through multiple channels. They exhibit less inertia and place more weight on best responding to the opponents they lost to and to the overall action distribution. The performance of their decisions are less sensitive to other

Table 8: Model of win rate of selected action (including decision weights and interactions)

Variables interacted	Main effect	# of games (sq. root)	Empirical win rate	Theoretical win rate	Skill	Experience	Time since rule change
Main effect	———	0.0076	−0.1376***	0.2683***	0.8313***	0.0066	0.0041
		(0.0048)	(0.0291)	(0.0067)	(0.0322)	(0.0115)	(0.0141)
Own previous action	−0.1083***	0.0119***	0.2021***	0.6139***	−0.6393***	−0.005	−0.0111
	(0.0092)	(0.0028)	(0.0173)	(0.0044)	(0.0165)	(0.0064)	(0.0079)
Action of opponents won against	−0.3420***	0.0811***	2.1579***	0.0467***	0.0819***	0.0362***	−0.0109†
	(0.0081)	(0.0030)	(0.0400)	(0.0032)	(0.0113)	(0.0047)	(0.0065)
Action of opponents lost to	−0.1378***	0.0544***	−0.811***	0.0131***	0.0872***	−0.0068	0.0050
	(0.007)	(0.0028)	(0.0308)	(0.0030)	(0.0101)	(0.0045)	(0.0060)
Population action distribution	−0.0068	0.0048	0.0113	−0.0089***	−0.0159	0.0304***	−0.0725***
	(0.0061)	(0.0033)	(0.0195)	(0.0026)	(0.0109)	(0.0045)	(0.0057)
Best response to opponents won against	−0.0155	0.1912***	0.8967***	0.0125**	−0.3103***	0.0327***	0.0229*
	(0.0101)	(0.0053)	(0.0451)	(0.0042)	(0.0167)	(0.0069)	(0.0096)
Best response to opponents lost to	0.0790***	0.1713***	−0.1174**	−0.0055	−0.1330***	0.0121†	0.0124
	(0.0096)	(0.0053)	(0.0382)	(0.0044)	(0.0164)	(0.0071)	(0.0102)
Best response to population action distribution	0.0339***	−0.0046*	0.0275*	0.0004	−0.0096	0.0014	−0.0700***
	(0.0036)	(0.0022)	(0.0125)	(0.0017)	(0.0065)	(0.0028)	(0.0043)
Action embedding deviation	−1.2211***	−0.036***	−0.0693†	−0.2562***	0.7986***	−0.0773***	0.0529**
	(0.0190)	(0.0069)	(0.0399)	(0.0080)	(0.0336)	(0.0140)	(0.0171)
Observations	1,298,271						
Fixed effects	32,584						
Overall R^2	35.40%						
Partial R^2 of all regressors	17.34%						
Partial R^2 of interactions	11.79%						

Note: ***: $p < 0.001$; **: $p < 0.01$; *: $p < 0.05$; †: $p < 0.1$. Standard errors are clustered at the player level but do not account for first-stage uncertainty from estimation of the neural network model and estimation of observation-level decision weights. For readability, the dependent variable and "theoretical win rate" are both multiplied by 100 (i.e., they are on percentage point scale). For the column (row) labeled "Main Effect," the coefficient entry gives the main effect of the corresponding row (column). The row variables correspond to estimated weights $\hat{\beta}_{i,t,p}$ placed on each decision rule for the given observation, except for "action embedding deviation" which corresponds to the Euclidean distance in the embedding space of the selected action from the value predicted by the estimated decision weights. All column variables (empirical win rate, number of games played, etc.) are centered before interaction, so the main effects for the row variables can be interpreted as the coefficients for those variables at the average levels of the column variables.

factors such as action error and imitating opponents, perhaps because they are able to identify aspects of opponent actions and deviations from the considered decision rules that are normatively well-adapted, or because they already place near-optimal weight on these variables.

Comparing these estimates to the estimates of the "how to change" model in Table 6 reveals striking differences: though the main effects of the normative model indicate that inertia and imitating opponents are bad for normative performance, while best responding to opponents lost to and to the population action distribution are good for normative performance, players on average place the vast majority of their decision weight on inertia and imitation and place very little (or even negative) weight on best responses. Additionally, comparing the signs of the interaction effects between the two tables, we have a mixed bag. For imitation of opponents lost to, the significant interactions are all in the same direction, indicating that players place more/less weight on this decision rule when it is more/less normatively advantageous to do so. However, the interaction effects for other decision rules often have opposite signs between the two models, indicating no consistent pattern of players placing more weight on decision rules when they are normatively better. Thus, while the comparison to Table 4 indicates that players are adaptive in deciding whether to change actions, they do not appear adaptive in choosing which heuristics/decision rules to rely on in selecting actions.

7.2 Managerial Implications of Player Behavior

The above analyses show that players make noisy and normatively suboptimal decisions. Though this finding is interesting from an academic perspective in terms of understanding how humans make decisions in complex competitive environments, it is also important to consider whether this normative suboptimality is managerially consequential for the platform hosting the competition. In particular, if players are motivated and kept engaged by winning, then making suboptimal decisions may lead them to become discouraged and less likely to continue playing the game. Conversely, if players are not motivated by winning but rather some other subjective notion of fun or engagement, then the normative optimality of player decisions should be largely inconsequential for the game platform.

To understand whether player decision-making is consequential for engagement, I perform a supplementary analysis wherein I observe players in the week before and after a rule change (since this is an external shock which essentially forces the player to change actions) and analyze how their performance affects their likelihood to continue playing the game. Specifically, for each of the four rule changes in the data, I look at players who played at least one game in the last week of the previous ruleset and one game in the first week of the new ruleset and model whether they continue to play at least one game under the new ruleset after the first week. I pool across the four rule changes and include fixed effects for the ruleset in effect. Player fixed effects are not feasible here, since there are few repeat observations, particularly for players who do not continue playing. By definition, if the dependent variable for a player is zero for a given ruleset, they cannot appear in the next ruleset. I control for the number of games played in the last/first week of the previous/new rulesets respectively as well as player skill.

Table 9: Model of whether players continue playing after a rule change

Variable	Coefficient (SE)
Number of games in previous ruleset (square root)	0.0055 (0.0087)
Number of games in new ruleset (square root)	0.2127 (0.0102)***
Empirical win rate in previous ruleset	0.0252 (0.1628)
Change in empirical win rate from previous ruleset (+)	−0.0316 (0.1774)
Change in empirical win rate from previous ruleset (−)	0.6803 (0.1495)***
Skill in new ruleset	0.2593 (0.0455)***
Observations	18,829
Fixed effects	4
Overall McFadden R^2	7.67%
Partial McFadden R^2 of regressors	7.56%

Note: ***: $p < 0.001$; **: $p < 0.01$; *: $p < 0.05$; †: $p < 0.1$. Variables for the previous (new) ruleset are calculated over the last (first) week of the ruleset, respectively. The relationship with change in empirical win rate is modeled as piecewise linear with separate coefficients for the negative and positive domains.

I specifically consider whether the change in empirical win rate from one ruleset to the next affects the player's likelihood of continuing playing, modeling the response as a piecewise linear function with a slope change at 0 to allow for loss aversion, i.e., asymmetric reactions to whether the player's performance became better or worse after the change (Kahneman and Tversky, 1979; Hardie et al., 1993). The results are given in Table 9. A key finding is that players' empirical win rates, both in the old and new ruleset, do not affect their likelihood of continuing to play, *except* when their performance became worse. Thus, when players experience a sudden drop in performance under a new ruleset, they may become discouraged and quit playing.

While I used empirical win rate in this model, the same patterns hold when substituting in the theoretical win rate calculated from my neural network model. Additionally, while this model focuses on action changes externally imposed by rule changes, I find the same directional results in a longitudinal analysis of player engagement over time (inclusive of player fixed effects), showing that players are less likely to play after making an action change that resulted in a drop in performance. These additional analyses are included in Appendix D.3. This finding is qualitatively consistent with the findings of Huang et al. (2019), who find through a hidden Markov model that the engagement state of a player is sensitive to their in-game performance, especially when that player is currently highly engaged.

These results show that players' decision-making processes are not only of theoretical interest for researchers studying behavioral decision-making, but also are consequential for the game platform itself. Players are substantially more likely to become frustrated and quit playing after experiencing a drop in performance, and Figure 13 shows that such a drop can happen with high probability. Thus, the game platform may be able to improve player retention by providing tools that can assist players in their decision-making process, e.g., by highlighting player blind spots in decision-making and suggesting modifications that can address those blind spots. Such tools have precedent in other games such as chess, where online game platforms provide game analysis tools that show players where they have made mistakes in the past and give suggestions on improving performance in the

future.

Of course, such an intervention would need to be relatively "light touch," and there are interesting trade-offs to consider: in one extreme, a tool that tells players the exact optimal action at a given time period would presumably not be well-received, as it would remove the challenge from the action selection process (which is presumably a source of fun and entertaintment for some players) and would lead to fast convergence towards a small subset of actions. In the context of a game played for entertainment, presumably the aggregate variability in the action distribution reflected in Figures 3 and 12 is desirable, since it keeps the game interesting for players. Thus, a more "light touch" approach may be, for instance, data visualization and information aggregation tools that make it easier for players to process information when making a decision (e.g., a dashboard that summarizes what opposing actions a player recently won/lost against and what the strategic strengths/weaknesses of those actions are) without solving the decision problem for them. Designing an appropriately balanced tool that can help prevent players from making majorly suboptimal decisions, while not undermining the entertainment value of the game as a whole, is an interesting platform design question to consider.

8 Discussion

In this paper, I developed a machine learning framework enabling the behavioral analysis of agent decisions in a complex game by mapping the large, unstructured space of possible player actions to a latent representation that encodes the strategic attributes of actions in a smooth, linear manner. I achieved this by mapping the representations of two actions to the pairwise probability that one action wins over the other via a bilinear functional form that enforces smoothness and linearity. I demonstrated the method on a large, novel dataset of decisions made by players of a competitive video game. My analysis showed that players can perform model-based reasoning and incorporate outside information even when making decisions in this complex setting, but they tend to rely much more on simple model-free heuristics and exhibit a high degree of noise in their decisions. I further showed that this noise and reliance on simple heuristics leads to substantial normative suboptimality in player decisions. In turn, players are more likely to become discouraged and quit playing after a setback where they choose a worse action, indicating that the normative suboptimality of player decisions is consequential for the game platform. Thus, not only is player behavior in e-sports a compelling empirical context for academic study of decision-making in complex competitive environments, it is also of considerable managerial relevance for game-playing platforms that wish to retain engagement from platform participants.

Though I performed my behavioral analysis using reduced form models of player decisions, my representation learning model can also allow for estimation of more structured mathematical models of player cognition in complex games. For instance, classic behavioral game theory models such as EWA treat players as maintaining separate "attractions" towards each action, with the player selecting a discrete action based on those attractions via a probabilistic decision rule such as multinomial

logit (Camerer and Ho, 1999). In a setting with millions of possible actions, maintaining separate attractions for each action is both computationally impractical for a modeler and behaviorally implausible for a player. In such contexts, it is more behaviorally plausible (and more computationally tractable) that players instead perform a degree of "smoothing," identifying actions that have similar strategic properties and pooling attractions across similar actions to achieve a more parsimonious representation of attractions over the action space. The representations learned by my model perform such smoothing by construction. A smoothed modification of the EWA model could be estimated using my method by, for example, modeling player attractions as a smooth parametric function over the embedding space or using kernel smoothing, and then approximating player action selection as a choice over the continuous embedding space. This can then yield estimates of explicit behavioral parameters and allow an apples-to-apples comparison between findings in complex settings such as mine and more classic behavioral game theory settings.

Even though my implementation of the model was developed based on my empirical context of a two-player, symmetric, zero-sum game, this general methodological approach can be used to analyze other types of non-cooperative games. In my context, I constrained the weight matrix of the bilinear form to be antisymmetric to encode the symmetric and zero-sum nature of payoffs. Games with asymmetric payoffs can be modeled by removing the antisymmetry constraint, and asymmetric action spaces can be modeled by having separate embedding neural networks for each player. Games that are not zero sum can be modeled using a separate bilinear form for each player, while games with more than two players can be modeled via general multilinear forms. Thus, this modeling framework can easily be generalized to extract efficient representations of large action spaces for other types of competitive decisions in general.

Additionally, my approach can be extended to model more abstract types of actions and competitions such as those involving unstructured data like text or images. For instance, advertisers often design competing ad campaigns, such as political campaigns that try to find the right messaging to contrast against opposing campaigns and convince constituents to vote for their cause. Modeling games where the "action" is an ad copy, and the payoff is the sentiment or downstream behavior of the target audience, could lead to new behavioral insights into how advertisers take competitor actions into account in their creative process. Existing pretrained models of unstructured data like images could be fine-tuned based on my proposed bilinear functional form to extract representations with the linear structure desired for such analyses. Generally, my proposed method can pair with recent advances in pretrained machine learning methods to analyze data on competitions with action spaces that previously would have been too complex to model.

In sum, I have proposed a representation learning framework to enable the behavioral analysis of games with complex action spaces and payoff structures. I demonstrated the utility of this method using a novel dataset on player behavior in a competitive video game. This analysis revealed that players can counterfactually but tend to make noisy decisions and overrely on simple heuristics to an extent comparable to findings from previous literature using much simpler games. The reliance on simple heuristics and noise in decision-making leads to normatively suboptimal outcomes, which tend

to result in lower player engagement, suggesting possible platform interventions to improve player retention. My general methodological approach is adaptable to a broad range of competitive settings involving complex or unstructured action spaces, suggesting interesting substantive applications to explore in future work.

References

Barberis, N. C. (2013). Thirty years of prospect theory in economics: A review and assessment. *Journal of Economic Perspectives*, 27(1):173–196.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828.

Burnap, A., Hauser, J. R., and Timoshenko, A. (2023). Product aesthetic design: A machine learning augmentation. *Marketing Science*.

Camerer, C. and Ho, T.-H. (1999). Experience-weighted attraction learning in normal form games. *Econometrica*, 67(4):827–874.

Camerer, C. F. and Ho, T.-H. (2015). Behavioral game theory experiments and modeling. In *Handbook of Game Theory with Economic Applications*, volume 4, pages 517–573. Elsevier.

Camerer, C. F., Ho, T.-H., and Chong, J.-K. (2002). Sophisticated experience-weighted attraction learning and strategic teaching in repeated games. *Journal of Economic Theory*, 104(1):137–188.

Chu, W., Im, M., and Lee, E.-J. (2014). Investor expertise as mastery over mind: Regulating loss affect for superior investment performance. *Psychology & Marketing*, 31(5):321–334.

Coibion, O. and Gorodnichenko, Y. (2015). Information rigidity and the expectations formation process: A simple framework and new facts. *American Economic Review*, 105(8):2644–2678.

Dangauthier, P., Herbrich, R., Minka, T., and Graepel, T. (2007). TrueSkill through time: Revisiting the history of chess. *Advances in Neural Information Processing Systems*, 20.

Dew, R., Ansari, A., and Toubia, O. (2022). Letting logos speak: Leveraging multiview representation learning for data-driven branding and logo design. *Marketing Science*, 41(2):401–425.

Elo, A. (1978). *The Rating of Chess Players, Past and Present*. Batsford chess books. Batsford.

Gabaix, X. (2019). Behavioral inattention. In *Handbook of Behavioral Economics: Applications and Foundations 1*, volume 2, pages 261–343. Elsevier.

Gathergood, J., Mahoney, N., Stewart, N., and Weber, J. (2019). How do individuals repay their debt? The balance-matching heuristic. *American Economic Review*, 109(3):844–75.

Gobet, F. and Simon, H. A. (1998). Expert chess memory: Revisiting the chunking hypothesis. *Memory*, 6(3):225–255.

Hardie, B. G., Johnson, E. J., and Fader, P. S. (1993). Modeling loss aversion and reference dependence effects on brand choice. *Marketing Science*, 12(4):378–394.

Haviv, A., Huang, Y., and Li, N. (2020). Intertemporal demand spillover effects on video game platforms. *Management Science*, 66(10):4788–4807.

Herbrich, R., Minka, T., and Graepel, T. (2006). TrueSkill™: a Bayesian skill rating system. *Advances in Neural Information Processing Systems*, 19.

Ho, T. H., Camerer, C. F., and Chong, J.-K. (2007). Self-tuning experience weighted attraction learning in games. *Journal of Economic Theory*, 133(1):177–198.

Hooshyar, D., Yousefi, M., and Lim, H. (2018). Data-driven approaches to game player modeling: a systematic literature review. *ACM Computing Surveys (CSUR)*, 50(6):1–19.

Houser, D., Keane, M., and McCabe, K. (2004). Behavior in a dynamic decision problem: An analysis of experimental evidence using a bayesian type classification algorithm. *Econometrica*, 72(3):781–822.

Howard, G. (2021). A check for rational inattention. Technical report, University Library of Munich, Germany.

Huang, Y., Jasin, S., and Manchanda, P. (2019). "Level up": Leveraging skill and engagement to maximize player game-play in online video games. *Information Systems Research*, 30(3):927–947.

Ishihara, M. and Ching, A. T. (2019). Dynamic demand for new and used durable goods without physical depreciation: The case of Japanese video games. *Marketing Science*, 38(3):392–416.

Joo, J. (2022). Express: Rational inattention as an empirical framework for discrete choice and consumer-welfare evaluation. *Journal of Marketing Research*, page 00222437221110173.

Kahneman, D. and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–292.

Khaw, M. W., Stevens, L., and Woodford, M. (2017). Discrete adjustment to a changing environment: Experimental evidence. *Journal of Monetary Economics*, 91:88–103.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Liu, J. and Ansari, A. (2020). Understanding consumer dynamic decision making under competing loyalty programs. *Journal of Marketing Research*, 57(3):422–444.

Matějka, F. and McKay, A. (2015). Rational inattention to discrete choices: A new foundation for the multinomial logit model. *American Economic Review*, 105(1):272–298.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In Zarembka, P., editor, *Frontiers in Econometrics*, pages 105–142. Academic Press.

McKelvey, R. D. and Palfrey, T. R. (1995). Quantal response equilibria for normal form games. *Games and Economic Behavior*, 10(1):6–38.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Nash, J. (1951). Non-cooperative games. *Annals of Mathematics*, 54(2):286–295.

Nevskaya, Y. and Albuquerque, P. (2019). How should firms manage excessive product use? A continuous-time demand model to test reward schedules, notifications, and time limits. *Journal of Marketing Research*, 56(3):379–400.

Ruiz, F. J., Athey, S., and Blei, D. M. (2020). Shopper: A probabilistic model of consumer choice with substitutes and complements. *The Annals of Applied Statistics*, 14(1):1–27.

Samuelson, L. (1998). *Evolutionary Games and Equilibrium Selection*, volume 1. MIT press.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Simonov, A., Ursu, R., and Zheng, C. (2022). Suspense and surprise in media product design: Evidence from twitch.tv.

Sims, C. A. (2003). Implications of rational inattention. *Journal of Monetary Economics*, 50(3):665–690.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R. R., and Smola, A. J. (2017). Deep sets. *Advances in Neural Information Processing Systems*, 30.